Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1470

Large language models for porous materials: from text mining to autonomous laboratory

Seunghee Han, Taeun Bae, Junho Kim, Younghun Kim and Jihan Kim *

Porous materials such as metal–organic frameworks (MOFs), covalent organic frameworks (COFs), zeolites, and porous carbons play central roles in gas storage, separation, catalysis, and environmental technologies. However, their design and discovery remain resource-intensive, relying heavily on expert intuition and fragmented knowledge distributed across the literature. Recent advances in large language models (LLMs) present new opportunities to accelerate these workflows by integrating scientific text mining, domain reasoning, and experimental planning. In this review, we outline the emerging role of LLMs across the porous materials research ecosystem. We first introduce the foundations of LLMs, followed by a discussion of NLP-based text mining for literature analysis. We then examine LLM adaptation including prompt engineering and fine-tuning, and autonomous research systems from human-in-the-loop to self-driving laboratories. For each domain, we summarize how LLM architectures are integrated with research systems, highlighting their applications, advantages, and limitations. Additionally, we discuss the current challenges of applying LLMs to porous materials, trade-offs between prompt engineering and fine-tuning, the influence of generation parameters such as temperature, and safety considerations in autonomous laboratory systems. Finally, we expect LLMs to advance toward multimodal reasoning, tighter integration with structured knowledge bases, and safer autonomous experimental workflows. Together, these developments suggest emerging LLM-driven paradigms that could transform the conceptualization, design, and synthesis of porous materials.

Received 23rd December 2025
Accepted 29th March 2026

DOI: 10.1039/d5dd00578g

rsc.li/digitaldiscovery

1. Introduction

Porous materials, including metal–organic frameworks (MOFs),[1] covalent organic frameworks (COFs),[2] porous coordination polymers,[3] zeolites,[4] and porous carbons,[5] have emerged as increasingly important materials systems owing to their intrinsic porosity and chemically tunable functionality. Their unique pore environments enable selective adsorption, ion transport, catalysis, and molecular recognition, resulting in applications across carbon capture, gas separation, sensing, drug delivery, and heterogeneous catalysis.[6–14] Despite continuous advances in synthesis strategies, computational tools, and high-throughput simulations, the design and discovery of porous materials remain highly resource-intensive.[15] Traditional workflows rely on expert intuition, iterative trial-and-error experimentation, and fragmented knowledge dispersed across publications, databases, and laboratory experience. As a result, the process of transforming a conceptual material hypothesis into a validated structure often requires months to years and is rarely reproducible at scale.[16]

While traditional ML models excel at numerical property prediction,[17,18] they do not directly leverage the vast amount of textual knowledge embedded in the literature. Recent advances in natural language processing (NLP) have enabled extraction of synthesis procedures, properties, and reaction conditions from unstructured scientific text, forming the basis for text-driven knowledge discovery.[19,20] More recently, large language models (LLMs) have expanded this direction by enabling not only information extraction but reasoning, summarization, decision-making, and multi-step workflow design.[21] Unlike previous approaches limited to quantitative structure–property mapping, LLMs enable integration of unstructured scientific knowledge with formal data representations. This capability expands their role from simple predictive modeling to hypothesis generation and knowledge synthesis in materials discovery.

These methodological advances are now beginning to reshape porous materials research specifically, where LLMs have been applied to tasks ranging from text mining to autonomous experimentation. Notably, most reported applications to date have focused on MOF systems, with more limited exploration of COFs, porous carbons, and zeolites, reflecting the current distribution of available studies. Early applications of LLMs in porous materials primarily focused on text mining, enabling the construction of synthesis databases

Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea. E-mail: jihankim@kaist.ac.kr

and automated entity extraction.[22] As the field progressed, LLMs increasingly shifted toward generative and decision-making roles, including inverse design, synthesis strategy recommendation, RAG-assisted reasoning, and multi-agent orchestration.[23] More recent work integrates LLMs with simulation tools and robotic experimentation, demonstrating early examples of closed-loop discovery pipelines and autonomous laboratory systems.[24] Collectively, these developments reflect a shift from passive language comprehension toward active participation in scientific reasoning and experimental planning.[25]

In this work, we provide a structured overview of how LLMs are being used across the porous materials research ecosystem. We first introduce core concepts and reasoning frameworks including the evolution of LLMs, prompt engineering,[26] chain-of-thought reasoning (CoT),[27] and retrieval augmented generation (RAG).[28] We then examine three major application domains: (1) natural language processing (NLP) for text mining, (2) LLM adaptation with prompt engineering, and fine-tuning and (3) autonomous systems progressing from human-in-the-loop frameworks toward self-driving laboratories (Fig. 1). Furthermore, we discuss current limitations, the trade-offs between prompt engineering and fine-tuning approaches, the role of temperature in the reliability of LLM-driven workflows, and safety considerations in autonomous laboratory systems. Finally, we outline future research directions. Together, these perspectives position LLMs not merely as computational tools but as emerging cognitive systems capable of connecting language, domain reasoning, and experimental execution. As multimodal modeling, structured knowledge representation, and autonomous experimentation advance, LLM-enabled workflows are expected to play an increasingly central role in how porous materials are conceptualized, designed, and synthesized.

2. Large Language Models (LLMs)

LLMs represent a rapidly advancing class of artificial intelligence (AI) systems built on massive corpora of text and code. Early generations of language models were based on recurrent neural networks (RNNs), which processed sequences step-by-step and therefore struggled to capture long-range dependencies. These models were widely applied across a range of scientific contexts, including next-word prediction, word translation, and sentiment classification. However, their strictly sequential nature made it difficult to retain information from earlier parts of a sentence when interpreting later content. The introduction of the Transformer architecture marked a major paradigm shift from this limitation, enabling parallel processing and effective modelling of global context.[29]

This shift opened the door for large-scale models and achieved their rapid expansion into different regimes,[30–32] where their ability to interpret, predict, and generate properties offers significant advantages. A key reason for this advancement is the Transformer's self-attention mechanism, which allows the model to examine all tokens in a sentence simultaneously rather than sequentially.[29] Through attention scores, Transformers capture long-range context, resolve complex dependencies, and represent nuanced relationships that earlier architectures struggled to encode. In particular, the transformer's capacity to recognize and encode complex linguistic patterns provides the foundation for how LLMs internalize scientific knowledge from literature and capture the relationships inherent in sequential, symbolic, and domain-specific data.

These Transformer-based models are pre-trained on massive amounts of text to learn linguistic structure and contextual relationships. After broad pretraining, they can be adapted to specialized downstream tasks through several complementary strategies. One such method is fine-tuning, in which a pre-trained model is trained further using curated literature and datasets relevant to specific scientific objectives.[33–35] Through this refinement, the model transforms a general-purpose architecture into a task-specialized assistant.

As model scale increased with systems such as GPT-2 and GPT-3, researchers discovered that LLMs could perform new tasks simply by conditioning them with natural-language instructions. This instruction-following capability is known as prompt engineering, a method that guides model behavior through carefully structured prompts without modifying model parameters. Because large pretrained models already encode broad scientific priors, prompt engineering often achieves strong performance without the computational cost of fine-tuning.

Complementing prompt-based control are advanced reasoning strategies designed to make the model's internal logic more transparent. Chain-of-Thought (CoT) prompting encourages models to articulate intermediate reasoning steps,[27] while Chain-of-Verification (CoV) instructs them to reevaluate their own conclusions by generating and answering verification questions to correct potential errors.[36] However, a fundamental limitation remains: these internal strategies fail when intermediate reasoning requires specific domain knowledge not reliably captured during pre-training. In chemistry, this often leads to logical hallucinations where the model generates plausible-sounding but scientifically flawed argumentation for niche reaction mechanisms or complex structural relationships. Building on these methods, the ReAct framework integrates explicit reasoning with action-taking by allowing the model to alternate between thinking and acting, which not only improves task performance but also provides more grounded intermediate reasoning steps in observable outcomes. Whereas CoT emphasizes internal reasoning transparency, ReAct introduces an explicit decision layer that links reasoning to tool use or external interaction. In parallel, the MRKL (Modular Reasoning, Knowledge, and Language) system addresses which module performs the task by routing queries to specialized expert modules or external tools, enabling more structured and reliable reasoning pipelines. The practical utility of these reasoning frameworks is illustrated in recent porous materials research. For instance, ReAct-style strategies enable iterative refinement of experimental workflows through interaction with simulated or real-world feedback, whereas MRKL-inspired architectures facilitate modular routing of queries to

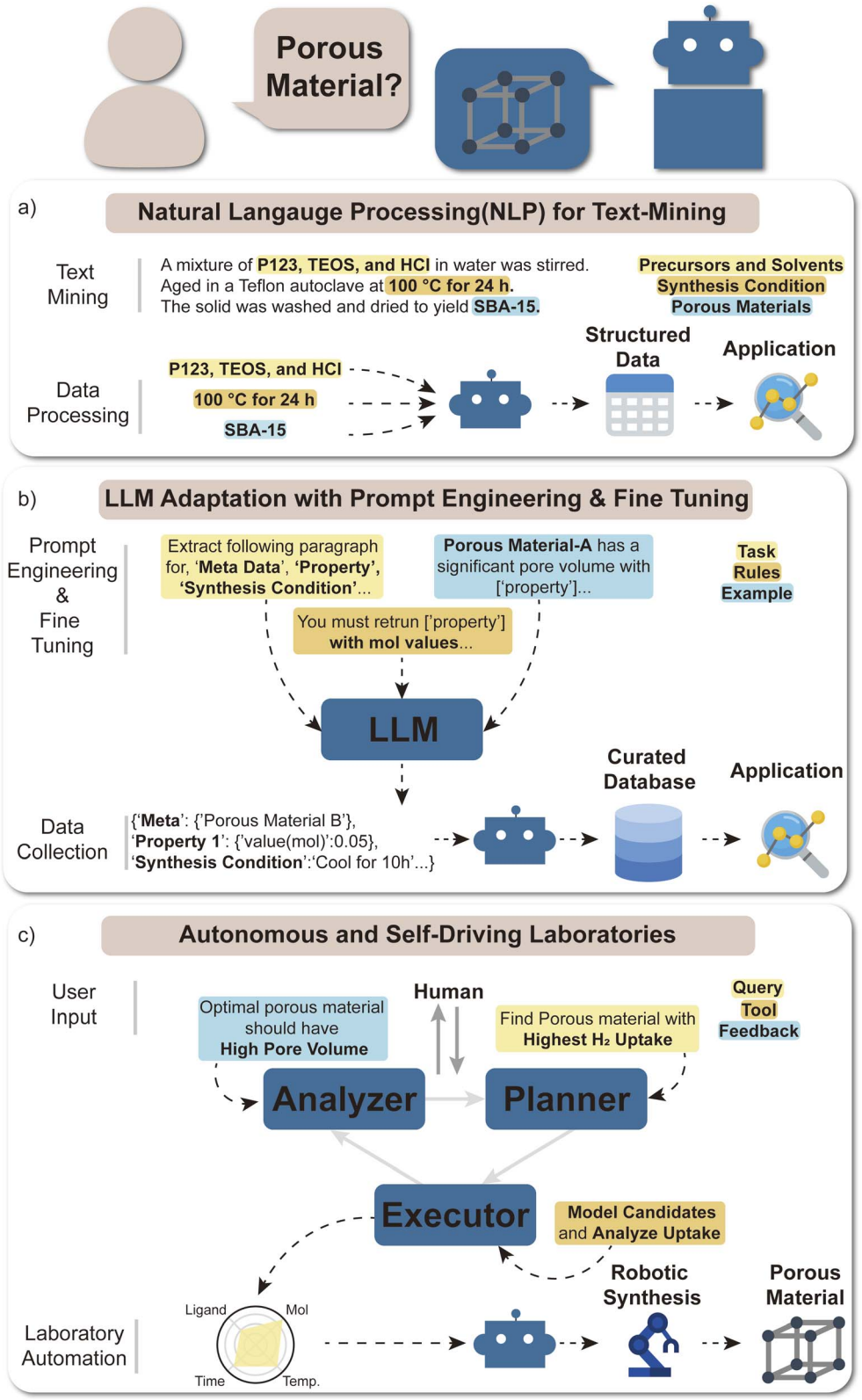

Fig. 1 Overall scheme illustrating how language-based methods have evolved within porous materials research: (a) natural language processing (NLP) for text mining of synthesis and property information, (b) the adaptation of LLMs through prompt engineering and fine-tuning to generate structured knowledge, and (c) autonomous systems covering the progression from human-in-the-loop operation to increasingly self-driving laboratory capabilities.

specialized computational or database tools. Rather than relying solely on internal parametric reasoning, such tool-integrated approaches allow intermediate steps to be supported by external calculations and structured operations. Importantly, the suitability of each framework remains inherently task-dependent. ReAct-style approaches are particularly useful when iterative interaction with tools or feedback is required to guide intermediate decisions, whereas MRKL-style architectures are advantageous when problem solving benefits from explicit decomposition into specialized computational or database modules. Detailed implementations of these strategies in porous materials systems are discussed in later sections. More broadly, these strategies are especially valuable in scientific research, where interpreting certain phenomena or structure–property relationships require multi-step argumentation.[37–39]

In the context of LLM-enabled scientific workflows, the terms verification and validation are often used interchangeably but represent conceptually distinct processes. In this review, verification refers to internal consistency checks within computational workflows, including schema enforcement, reasoning self-checks (*e.g.*, chain-of-verification), and cross-database grounding. Validation, in contrast, denotes the assessment of scientific correctness and real-world applicability, often requiring expert review or experimental confirmation. While earlier sections may have used these terms more loosely, we here adopt a clear distinction between internal verification procedures and external scientific validation, particularly as workflows progress toward autonomous systems.

Despite their capabilities, LLMs are prone to hallucinations, a phenomenon in which the model confidently generates statements that appear plausible but are factually incorrect or unsupported by data. This challenge becomes especially serious in scientific contexts, where experimental conditions or structural descriptors may appear only in isolated publications or supplementary data files.[40] Retrieval augmented generation (RAG) was developed in direct response to these limitations.[28] By grounding its answers in verified sources, RAG significantly reduces hallucination and improves factual consistency in chemical reasoning. RAG's performance is strictly governed by boundary conditions, such as the quality of the knowledge base and the retriever's precision with domain-specific terminology. Common failure modes in scientific workflow, which often compromise this precision, include retrieval errors, where irrelevant documents are fetched, and integration errors, where the model defaults to internal priors despite the provided context. Addressing these retrieval and integration failures is therefore essential for maintaining faithfulness and robustness in knowledge-intensive chemical discovery.

Together, these techniques form a coherent ecosystem that governs how LLMs acquire domain knowledge, interpret scientific information, and maintain reliability. Given the complex, multiscale datasets, specialized terminology, and highly fragmented reporting across the porous materials community, such reasoning-oriented and retrieval-grounded approaches are increasingly essential.

The rapid evolution of LLMs has produced a diverse landscape of powerful models, including GPT-4, Claude, Llama, Nemotron, Mistral, ChatGLM, Falcon, and DeepSeek.[41–49] These systems differ significantly in accessibility: GPT-4 and Claude are proprietary commercial models available through paid APIs, whereas Llama (Meta), Mistral, ChatGLM2-6B, Falcon-7B, Nemotron (NVIDIA), and DeepSeek offer open-source or openly licensed weights that support local deployment and customization. In addition, models like OpenAI's GPT-4V and Qwen-VL introduce multimodal capabilities, enabling the model to interpret and reason over both images and text, thereby broadening its utility across scientific and technical applications.[50] Their increased multimodal capacity and stronger reasoning abilities have made them particularly suitable for scientific discovery, where text, numerical descriptors, and experimental data must be interpreted together. For instance, these models can perform visual reasoning on SEM or TEM micrographs to identify morphological defects or pore-filling patterns that are difficult to encode into traditional descriptors. By interpreting these visual cues, LLMs can provide closed-loop feedback for synthesis planning, autonomously suggesting adjustments to reaction conditions based on the visual confirmation of a product's structural integrity. This transition from passive image recognition to active visual reasoning represents a functional capability that traditional ML predictors lack, positioning multimodal LLMs as essential decision-makers in autonomous laboratory environments. As these advanced architectures have matured, they have accelerated a broader methodological shift from traditional text mining pipelines to multimodal, retrieval-grounded, and reasoning-aware computational workflows.[33–35] This transition reflects a fundamental change: LLMs are no longer tools for passive information extraction but active engines that support hypothesis generation, decision-making, and integrated scientific analysis.

In the following section, we explore how language models have been applied in materials science, followed by an in-depth examination of their applications in porous materials. An overview of representative studies is summarized in Table 1, which includes information such as publication year, material system, model, method, key findings, and limitations. To ensure transparency, the studies in Table 1 were selected *via* Google Scholar using keywords including 'LLM,' 'text-mining,' 'porous material,' and 'self-driving lab.' This selection represents key integration cases, with individual references provided for further detail. In our effort to provide a structured overview in Table 1, we have categorized the limitations of existing studies into standardized themes such as data dependency, limited generalization, prompt sensitivity, reliability and hallucination risks, human-in-the-loop dependence, and automation, scalability constraints, or scope limitation. However, it is important to note that establishing a perfectly uniform categorization remains challenging at this stage. This difficulty arises primarily from the high degree of heterogeneity in current research, where the lack of standardized reporting for prompt configurations and the

Table 1 Summary of LLM based systems applied to porous materials (searched through Dec 2025). Each entry reports the system name, year, target material, model, adaptation strategy, key findings, and limitations

System	Year	Material	LLM model	Prompt engineering or finetuning	Key findings	Limitation
CCA[51]	2023	MOF	GPT-3.5-Turbo, GPT-4	Prompt engineering	Prompt-only LLM-based extraction of MOF synthesis data with high accuracy	Data dependency: only for well-formatted text
Paragraph2MOFInfo[33]	2024	MOF	GPT-3.5-Turbo, GPT-4, Mistral-7B, Llama-2, Llama-3, T5, BART	Prompt engineering, finetuning	Fine-tuned LLM-based extraction of MOF synthesis and property information	Hallucination & reliability risk: limited coverage of complex extraction targets
GPT-4V Image Mining[52]	2024	MOF	GPT-4V	Prompt engineering	Multimodal LLM-based mining of MOF characterization data from figures	Prompt sensitivity: strong sensitivity to prompt design and category definition
L2M3 (ref. 53)	2025	MOF	GPT-3.5-Turbo, GPT-4	Prompt engineering, finetuning	Data-driven MOF synthesis condition recommendation using LLM-extracted databases	Hallucination & reliability risk: extraction accuracy constrained by LLM performance
Porous Carbon Mining[54]	2025	Porous carbon	ChatGPT 4.0 API	Prompt engineering	Integrated LLM-AutoML framework for inverse design of porous carbons	Hallucination & reliability risk: requirement for post-extraction data curation and verification
SYN-COF[55]	2025	COF	Deepseek-R1	Prompt engineering	LLM-based COF synthesis extraction and ML prediction with experimental validation	Limited generalization: limited to literature-rich dual-monomer solvothermal COFs
NERRE Extractor[56]	2024	MOF	GPT-3, Llama-2	Finetuning	Fine-tuned LLM extraction of hierarchical and relational scientific information	Hallucination & reliability risk: formatting inconsistency
Eunomia[57]	2024	MOF	GPT-4	Prompt engineering	ReAct agent-based LLM system for materials data extraction	Prompt sensitivity: reliance on prompts and auxiliary tools
MOF-LLM Performance Benchmark[58]	2024	MOF	Llama2-7B, ChatGLM2-6B, Vicuna-7B, Falcon-7B, Mistral-7B, Marcoroni-7B, Llama2-13B, Vicuna-13B	Prompt engineering	Systematic benchmarking of open-source LLMs for diverse MOF research tasks	Hallucination & reliability risk: insufficient MOF-specific domain knowledge without fine-tuning
RetChemQA[59]	2024	MOF	GPT-4-Turbo	Prompt engineering	Large-scale single- and multi-hop QA benchmark for reticular chemistry	Scope limitation: restriction to literature-grounded question answering rather than material generation
LLM-Based Hydrophobicity Predictor[60]	2025	MOF	Gemini-1.5 Flash	Finetuning	Text-based prediction of MOF hydrophobicity using fine-tuned LLMs	Limited generalization: weak for unseen solvent- or ion-containing MOFs
MOF Linker Mutation Model[61]	2023	MOF	GPT-3.5-Turbo	Finetuning	Fine-tuned LLM generation of chemically valid MOF linker mutations	Human-in-the-loop dependency: need for human validation of chemical plausibility and synthetic feasibility
GPT-4 Reticular Chemist[62]	2023	MOF	GPT-4	Prompt engineering	LLM-human collaboration enabling the design of four new isoreticular MOFs	Hallucination & reliability risk: limited capability of GPT-4 in property assessment of MOFs requiring human expertise
MOFsyn agent[63]	2025	MOF	GPT-4o, Deepseek-V3, GLM-4-Flash-250414, Qwen2.5-MAX	Prompt engineering	Stepwise reduction strategy proposed for catalyst performance optimization using LLM	Human-in-the-loop dependency: reliance on manual experimentation
OSDA Design[64]	2025	Zeolite		Prompt engineering	OSDA distribution sampling with proposal	Hallucination & reliability risk: limited LLM

Table 1 (*Contd.*)

System	Year	Material	LLM model	Prompt engineering or finetuning	Key findings	Limitation
			GPT-4o, GPT-3.5-turbo, Llama 3.1, Llama 3.2, Nemotron-4		of new high-affinity candidates	capability in synthesizability assessment and synthesis pathway estimation for complex OSDAs
SciToolAgent[65]	2025	MOF	GPT-4o, OpenAI-o1, Qwen2.5-72B	GPT-4: prompt engineering/Qwen: finetuning using LoRA	State-of-the-art performance in scientific tool evaluation with large-scale tool orchestration and integrated safety checks	Limited automation & scalability: due to manual knowledge graph construction, and reliance on GPT-4o
dZiner[66]	2024	MOF	GPT-4o, Claude 3.5 Sonnet	Prompt engineering	LLM agent-driven inverse design from properties to structures	Data dependency: oversimplification of complex MOF
ChatMOF[23]	2024	MOF	GPT-4, GPT-3.5-turbo, GPT-3.5-turbo-16k	Prompt engineering	Autonomous MOF search, property prediction, and inverse design enabled by ReAct/MRKL architecture	Prompt sensitivity: token-length limitations, hallucination & reliability risk: occasional reasoning failures, and reduced generative diversity during MOF generation
ChatGPT Research Group[24]	2023	MOF, COF	GPT-4	Prompt engineering	Integration of 7 AI agents with bayesian optimization for efficient MOF and COF crystallinity optimization	Limited automation & scalability: requiring more advanced robotic platforms
MOFGen[67]	2025	MOF	GPT-4, Llama	Prompt engineering	Modular multi-agent framework combining LLMs, diffusion models, and QM agents with experimental realization of five AI-dreamt MOFs	Human-in-the-loop dependency: for exploring synthetic possibilities
Zn-HKUST-1 Green Synthesis[68]	2025	MOF	ChatGPT	Prompt engineering	Sustainable MOF synthesis optimization *via* LLM-based planning and high-throughput pipetting robots	Limited automation & scalability, human-in-the-loop dependency: human intervention required in experimental workflows

frequent use of non-public, proprietary datasets hinder direct cross-study comparisons.

2.1 Natural Language Processing (NLP) for text mining

Traditionally, materials researchers have acquired data through manual literature review and experimental investigation. However, the rapid growth of scientific publications and the expansion of chemical design space have rendered manual data collection increasingly inefficient. In response, natural language processing (NLP)-based text mining has emerged as a powerful paradigm for automatically converting unstructured textual information into structured, machine-readable data, enabling large-scale extraction of materials knowledge from the literature.[69,70]

At a fundamental level, NLP-based text mining pipelines in materials science can be decomposed into three core processes: text preprocessing, text representation, and information extraction, each of which incorporates a distinct set of computational techniques and models.

Text preprocessing focuses on converting raw, unstructured text into linguistically analyzable units. This step typically includes sentence segmentation, tokenization, part-of-speech tagging, and syntactic or dependency parsing, which together provide a structural foundation for downstream analysis. In materials-oriented workflows, preprocessing further involves the identification of chemically relevant text segments, such as synthesis-related paragraphs, often using lightweight classification models including logistic regression. Auxiliary NLP tools such as ChemicalTagger[71] and dependency parsers such as Stanza[72] are also employed at this stage to recognize experimental actions and syntactic relationships between entities described in scientific text.

Text representation transforms preprocessed linguistic units into numerical embeddings that can be consumed by machine-learning models. Early materials text mining studies relied on

static word embeddings such as Word2Vec,[73] whereas more recent approaches increasingly adopt contextual language models, most notably bidirectional encoder representations from transformers (BERT).[74] Domain-adapted variants, including PubMedBERT[75] further pretrained on materials science corpora, are often used to capture the specialized semantics of chemical terminology and experimental descriptions, providing high-quality representations for downstream extraction tasks.

Information extraction constitutes the central component of NLP-driven text mining, where structured knowledge is derived from text. Named entity recognition (NER) is widely used to identify materials entities such as chemical compounds, precursors, solvents, synthesis conditions, and properties, commonly implemented using sequence-labeling architectures including BiLSTM–CRF[76] networks or transformer-based encoders. To complement data-driven models, rule-based techniques such as regular expressions, keyword matching, domain-specific lexicons, and ontology-guided filters are extensively integrated to extract numerical values and synthesis descriptors that exhibit high linguistic variability. Beyond free-text content, table parsing algorithms are employed to extract property values and synthesis parameters reported in tabular formats. Chemistry-aware NLP frameworks such as ChemDataExtractor[77] integrates many of these extraction strategies with document structure analysis and chemical entity recognition, enabling scalable and automated information extraction across large and heterogeneous literature collections.

Collectively, these three processes form a unified pipeline in which statistical learning, deep neural networks, and rule-based heuristics are combined to transform unstructured materials literature into structured, machine-readable datasets. These components support a wide range of NLP applications across materials domains. As an illustrative example, Shetty *et al.* presented an NLP pipeline for polymer property extraction by developing MaterialsBERT, obtained by further pre-training PubMedBERT on 2.4 million materials science abstracts.[78] Applied to polymer literature, the system produced over 300 000 property records from 130 000 documents, demonstrating the scalability of NLP-driven data extraction in materials science. Kononova *et al.* developed one of the first large-scale NLP pipelines for inorganic synthesis extraction using a BiLSTM–CRF model, dependency parsing with neural networks, keyword-based matching, and a specialized material parser.[79] The system produced 19 488 solid-state reactions involving 13 009 targets and 1845 precursors, establishing the first automated large-scale synthesis-pathway dataset. These examples highlight how NLP has been applied in materials science, and in the following section, we take a closer look at its use in porous materials through several representative studies.

2.1.1 NLP-driven data extraction and database construction. NLP methods applied to porous materials encompass a broad range of tasks, including the extraction of quantitative descriptors and other well-structured information from the literature. Many of these approaches integrate rule-based text mining techniques with structural analysis to link experimentally reported data to computationally derived properties.

Tayfuroglu *et al.* employed a large-scale text and data mining (TDM) workflow to investigate H_2 uptake in MOFs.[80] A rule-based NLP pipeline incorporating tokenization and keyword-based extraction was used to collect surface area (SA) and pore volume (PV) information from 58 700 publications, resulting in SA values for 5975 MOFs and PV values for 7481 MOFs. The NLP approach achieved accuracies of 78% for SA and 82% for PV. In parallel, theoretical SA and PV values were computed for 72 000 structures in the Cambridge Structural Database (CSD)[81] using Zeo++.[82] These theoretical descriptors were integrated with experimentally extracted TDM values to estimate H_2 uptake, and the resulting predictions showed good agreement with grand canonical Monte Carlo (GCMC) simulations. Collectively, this study illustrates that rule-based NLP, when combined with structural modeling, offers a scalable and effective framework for data-driven evaluation of hydrogen storage performance in MOFs.

To expand beyond the extraction of numerical descriptors, a subsequent study incorporated chemistry-aware NLP toolkits and additional structural metadata to capture a broader range of synthesis-related information. Glasby *et al.* developed DigiMOF, an automatically generated database of MOF synthesis information obtained through large-scale text mining (Fig. 2a).[83] Using the chemistry-aware NLP toolkit ChemDataExtractor,[77] the workflow extracted key synthesis descriptors including synthesis methods, solvents, linkers, and metal precursors from 43 281 MOF-related publications, resulting in 15 501 unique MOFs and 52 680 synthesis-related property records. To enrich the dataset, additional metadata such as topological and geometric features were integrated using CrystalNets[84] and Zeo++,[82] enabling systematic connections between synthesis conditions and structural characteristics. The pipeline achieved a precision of approximately 77%, offering a robust, large-scale dataset suitable for data-driven studies of MOF synthesis. Overall, this work established a comprehensive digital infrastructure for the extraction and analysis of synthesis data in porous materials research.

Building on developments in MOF-focused data extraction, NLP-driven approaches have also been extended to other porous material families. Pan *et al.* developed ZeoSyn, a comprehensive zeolite synthesis dataset aimed at systematically mapping the large chemical space of zeolites.[85] To construct the dataset, the authors implemented an NLP-driven pipeline that integrates table parsing, named entity recognition, regular expressions, and domain-specific keyword matching to extract synthesis parameters such as gel compositions, reaction conditions, inorganic precursors, organic structure-directing agents (OSDAs), and product frameworks from 3096 journal articles. Following extensive manual verification, the resulting dataset comprised 23 961 synthesis routes covering 233 zeolite topologies and 921 unique OSDAs. To illustrate the analytical utility of the dataset, the authors applied SHapley Additive exPlanations (SHAP) to assess how specific synthesis parameters influence the likelihood of forming particular zeolite frameworks. Altogether, this work established a rigorously curated and interpretable resource that supports data-driven investigation and informed design of zeolite synthesis.

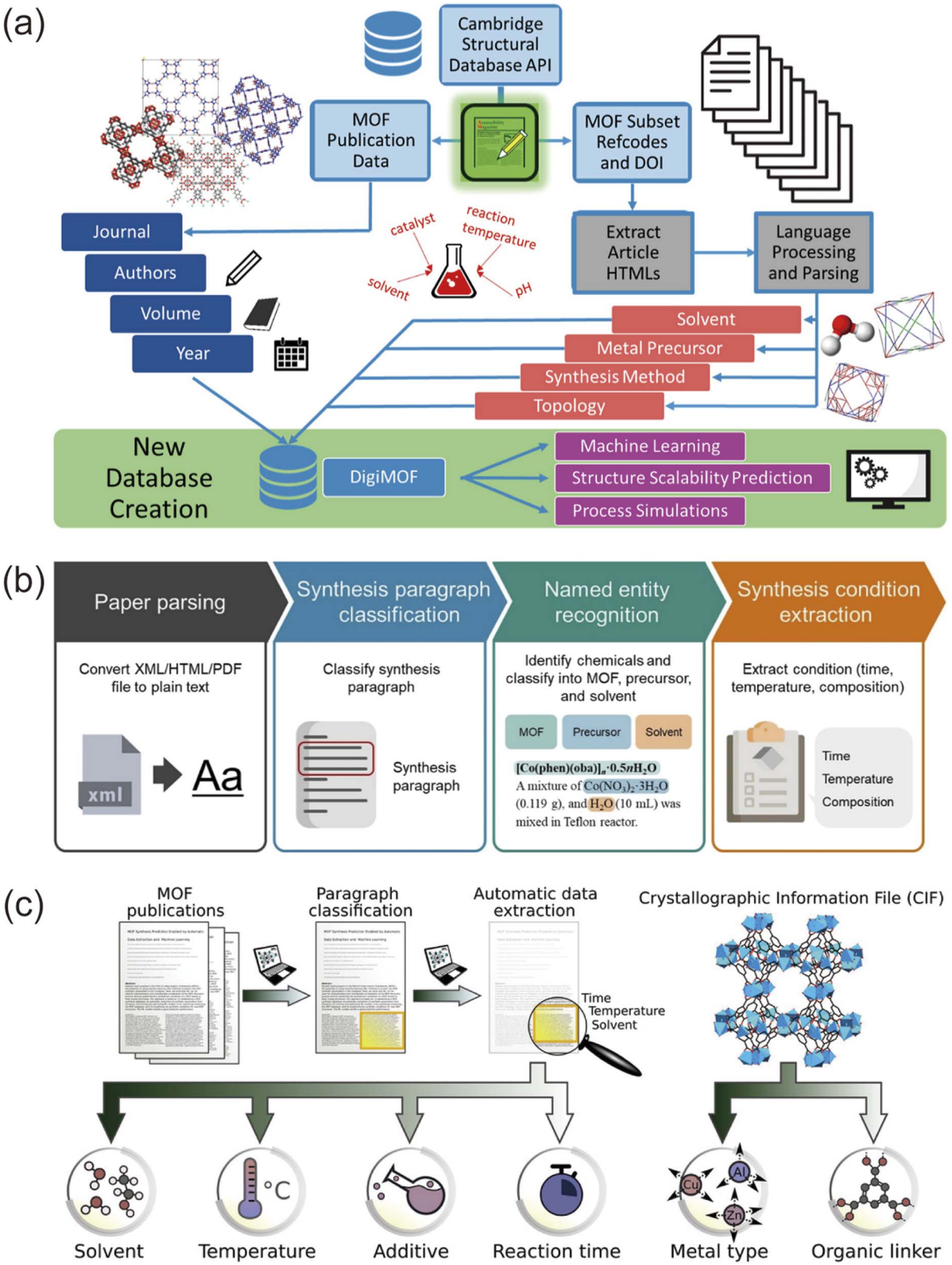

Fig. 2 NLP-enabled text mining pipelines for automated data extraction and ML-driven synthesis prediction in MOF (a) large-scale literature mining to extract structured MOF synthesis information. Adapted from ref. 82 and licensed under CC-BY 4.0. (b) Literature-derived databases supporting statistical analysis and crystallization outcome prediction. Adapted from ref. 20. Copyright 2022 American Chemical Society (c) ML models trained on NLP-extracted data to predict synthesis conditions directly from MOF structures. Adapted from ref. 85 and licensed under CC-BY 4.0.

Building on parameter-level synthesis datasets such as ZeoSyn, recent work has moved toward extracting complete experimental procedures from text. He *et al.* developed ZeoReader, an end-to-end information extraction framework for reconstructing structured zeolite synthesis steps directly from the literature.[86] Rather than extracting isolated synthesis parameters, ZeoReader models synthesis as event-level sequences composed of modular actions such as add, stir, and crystallize together with associated properties including temperature, duration, pressure, and materials. The framework consists of PDF parsing, a MatSciBERT-based paragraph classifier for identifying synthesis-relevant passages, and a two-stage event extraction model. Action detection is formulated as trigger classification, while property extraction is implemented using a BART encoder–decoder model in which predefined action-specific templates, such as "add material to container at temperature", are completed by filling the material, container, and temperature slots with text spans from the original sentence. To improve robustness in sentences containing multiple densely packed properties or bracketed quantities, the authors introduced contrastive learning by constructing correctly populated templates as positive samples and partially or incorrectly filled templates as negative samples. An Information Noise-Contrastive Estimation (InfoNCE) loss is used to bring representations of correct templates closer to the original instance and push incorrect ones apart, reducing omission and boundary errors during property extraction. The system achieved 94.06 percent accuracy for paragraph identification and an F_1 score of 74.99 percent for property extraction. The authors note limitations including the inability to process tables or multimodal content, dependence on a predefined synthesis schema, and difficulty handling unseen action types or cross-sentence dependencies. Overall, ZeoReader extends NLP-driven extraction in porous materials from parameter aggregation toward structured, step-level synthesis knowledge.

2.1.2 NLP-enabled applications. Collectively, these studies demonstrate that NLP-based text mining can produce structured and information-rich datasets for porous materials. Building on this foundation, subsequent research has explored how these literature-derived datasets can be utilized to support predictive modelling, synthesis planning, and other data-driven applications in porous materials research.

Park *et al.* developed a large-scale literature-mined database of MOF synthesis information to enable data-driven prediction of crystallization outcomes (Fig. 2b).[20] Their NLP pipeline combined logistic regression for classifying synthesis-relevant paragraphs, a BiLSTM–CRF[76] model for named entity recognition of MOF names, precursors, and solvents, and rule-based regular expressions for extracting reaction conditions, including temperature and time. Applying this workflow to 28 565 MOF-related publications yielded 46 701 synthesis records. Because unsuccessful synthesis attempts are rarely reported in the literature, the authors trained a positive-unlabeled learning model to predict MOF crystallinity, termed the crystal score. The model achieved a recall of 83% and successfully distinguished reported amorphous MOFs from their crystalline counterparts in multiple case studies. Overall, this work established a methodological basis for automated, data-driven analysis of MOF synthesis and demonstrated the broader potential of NLP-extracted literature data for predictive materials discovery.

Whereas this study focused on predicting crystallization outcomes from literature-derived data, further research has examined how NLP-extracted information can support the inference of synthesis parameters from structural inputs. Luo *et al.* developed a machine-learning framework that utilizes NLP-extracted synthesis data to predict MOF synthesis conditions directly from crystal structures, aiming to move beyond empirical, trial-and-error approaches toward data-driven synthesis design (Fig. 2c).[87] Using ChemicalTagger[71] for entity recognition and rule-based filtering, the authors extracted synthesis information such as metal source, linker, solvent, additive, temperature, and reaction time from publications linked to Computation-Ready Experimental MOF (CoRE MOF)[88] and CSD entries, constructing the SynMOF dataset containing 983 MOFs. Machine learning models were trained using molecular fingerprints of organic linkers together with metal identity and oxidation state to predict synthesis parameters. The resulting models achieved positive predictive performance, with meaningful R^2 values for temperature and reaction time prediction and top-three solvent selection accuracy exceeding 90% for single-solvent systems. These results suggest that NLP-derived synthesis data can be systematically integrated with structural information to support data-driven and predictive planning of MOF synthesis.

In addition to synthesis prediction, NLP-derived datasets have also been used to examine other material characteristics relevant to practical implementation, including stability. Nandy *et al.* developed MOFSimplify, a data-driven platform that integrates NLP-extracted experimental stability data with machine learning models to predict the robustness of MOFs.[89] Using ChemDataExtractor[77] for sentence tokenization and Stanza[72] for dependency parsing in combination with regular expressions, the authors mined over 5000 publications associated with the CoRE MOF,[88] extracting 2179 solvent removal stability labels and 3132 thermal decomposition temperatures. These data were combined with revised autocorrelation (RAC) descriptors,[90] which capture coordination chemistry, and Zeo++[82] geometric features representing pore topology. The resulting dataset was used to train artificial neural network models, achieving 76% accuracy for solvent removal stability classification and a mean absolute error of 47 °C for thermal stability prediction. By systematically linking literature-mined stability data to experimentally resolved MOF structures, the study demonstrates that NLP-based workflows can generate large scale, chemically meaningful stability datasets. The MOFSimplify web platform provides open access to these data and models, representing a significant step toward data-driven, automated prediction and design of stable MOFs.

Beyond stability analysis, NLP-extracted data have also been integrated with generative machine-learning models to explore inverse design tasks in porous materials. Jensen *et al.* demonstrated a data-driven framework for inverse design of OSDAs by applying NLP to the zeolite synthesis literature.[91] From more than 5000 reported synthesis routes from 1384 publications, the

authors identified relationships among 758 OSDAs and 205 zeolite frameworks. The OSDAs were further characterized using weighted holistic invariant molecular (WHIM) descriptors[92,93] to capture shape-matching effects with zeolite cavities. Using this literature-derived dataset, the authors trained a generative recurrent neural network (RNN) conditioned on zeolite topology and gel chemistry to propose new OSDA candidates. The model successfully regenerated known OSDAs and proposed new candidates for frameworks such as CHA and SFW. This study highlights how NLP-extracted synthesis knowledge, when integrated with molecular descriptors and generative machine learning, can enable inverse design in zeolite synthesis.

Although NLP-based text mining has enabled large-scale extraction of synthesis conditions, stability metrics, and structure–property relationships for porous materials, its performance remains constrained by incomplete literature access, heterogeneous reporting styles, and limited ability to capture complex experimental variables. As highlighted in recent quantitative analyses, variations in table structures, variable-based expressions, and inconsistent unit or identifier usage can substantially affect extraction performance.[94] These factors introduce noise, reduce generalizability, and limit the predictive accuracy of downstream machine learning models. Moreover, commonly reported exact-match accuracy metrics are sensitive to formatting differences, normalization procedures, and multiple valid chemical representations, and therefore should be interpreted in the context of their original evaluation protocols. The level of acceptable accuracy may also vary depending on the intended downstream application. Despite these challenges, existing studies demonstrate that rule-based and classical NLP workflows can compile chemically meaningful datasets at scale. Building on these foundations, recent advances in LLMs offer the potential to overcome many of these limitations by improving entity recognition, contextual understanding, and extraction of experimental information.

2.2 LLM adaptation with prompt engineering and fine-tuning

In recent years, LLMs have been actively applied across diverse domains of chemistry and materials science including porous materials. In literature-based data extraction, ChatExtract demonstrated that zero-shot conversational LLMs, guided by carefully engineered prompts and iterative questioning, can automatically construct high-quality materials property databases directly from scientific text and tables.[40] It successfully extracted structured datasets such as the critical cooling rates of metallic glasses and yield strengths of high-entropy alloys with precision and recall approaching 90%, highlighting the potential of LLMs in autonomous knowledge curation. At the level of scientific text representation, Choi and Lee showed that GPT-based LLMs can replace complex architectures for materials language processing by using prompt engineering, enabling high-performance zero/few-shot text classification, NER, and extractive QA on limited datasets.[95] For molecular and property discovery, LLM4SD (LLMs for Scientific Discovery) showcased how LLMs can be used not only as black-box predictors but also as interpretable reasoning tools.[96] The model extracted scientific rules from literature and learned new correlations from SMILES-based data, converting them into interpretable feature vectors that, when combined with traditional ML methods, outperformed graph neural networks on molecular property prediction tasks.

In inorganic synthesis, both Schrier *et al.*[97] and Kim *et al.*[98] fine-tuned GPT-based models for synthesis prediction tasks. Schrier *et al.* demonstrated that fine-tuned GPT-3.5/4 models can predict the synthesizability and suitable precursors of inorganic compounds, achieving performance comparable to specialized graph-based machine learning models. Similarly, Kim *et al.* employed a lightweight fine-tuning of GPT-4o mini to predict the synthesizability of inorganic crystal polymorphs, further introducing an explainable framework where LLM-generated natural-language reasoning revealed key compositional and structural determinants of synthetic feasibility.

In addition to text-only reasoning, recent studies have begun to benchmark the multimodal capabilities of LLMs. MaCBench benchmarked the ability of vision–language models (VLMs) to process images, tables, and spectra in chemistry and materials contexts.[99] While models like Claude 3.5 and GPT-4V showed strong performance in equipment recognition and simple data extraction, they still exhibited limitations in spatial reasoning and multi-step inference, which are essential for tasks such as spectral interpretation and crystal structure analysis.

Together, these studies demonstrate that LLMs are rapidly becoming integral components of chemical and materials research pipelines, supporting tasks ranging from literature-driven data extraction to interpretable molecular discovery, synthesis planning, and multimodal scientific reasoning.

2.2.1 LLM-assisted data mining and knowledge extraction in porous materials. The application of LLMs to porous materials has primarily focused on the automation of literature-derived data curation, where experimental synthesis conditions and material properties are converted into structured, machine-readable formats.

As an initial demonstration, Zheng *et al.*[51] developed the ChatGPT Chemistry Assistant (CCA) for mining MOF synthesis information from unstructured text. By employing domain-specific prompt templates (ChemPrompt Engineering), ChatGPT autonomously performed text filtering, paragraph classification, and synthesis-parameter summarization, which were previously implemented through manually coded NLP pipelines. The overall literature-to-database workflow implemented by the CCA, including human preselection, paragraph classification, and prompt-guided synthesis-condition summarization, is schematically illustrated in Fig. 3a. Across 228 representative MOF publications, the CCA extracted over 26 000 synthesis parameters (metal sources, organic linkers, solvents, temperatures), achieving 90–99% precision and recall. The extracted data further enabled a supervised machine-learning model to predict crystallization outcomes with accuracy exceeding 87%. This work demonstrated the feasibility of using general-purpose conversational LLMs to perform chemical information extraction at near-human reliability, providing

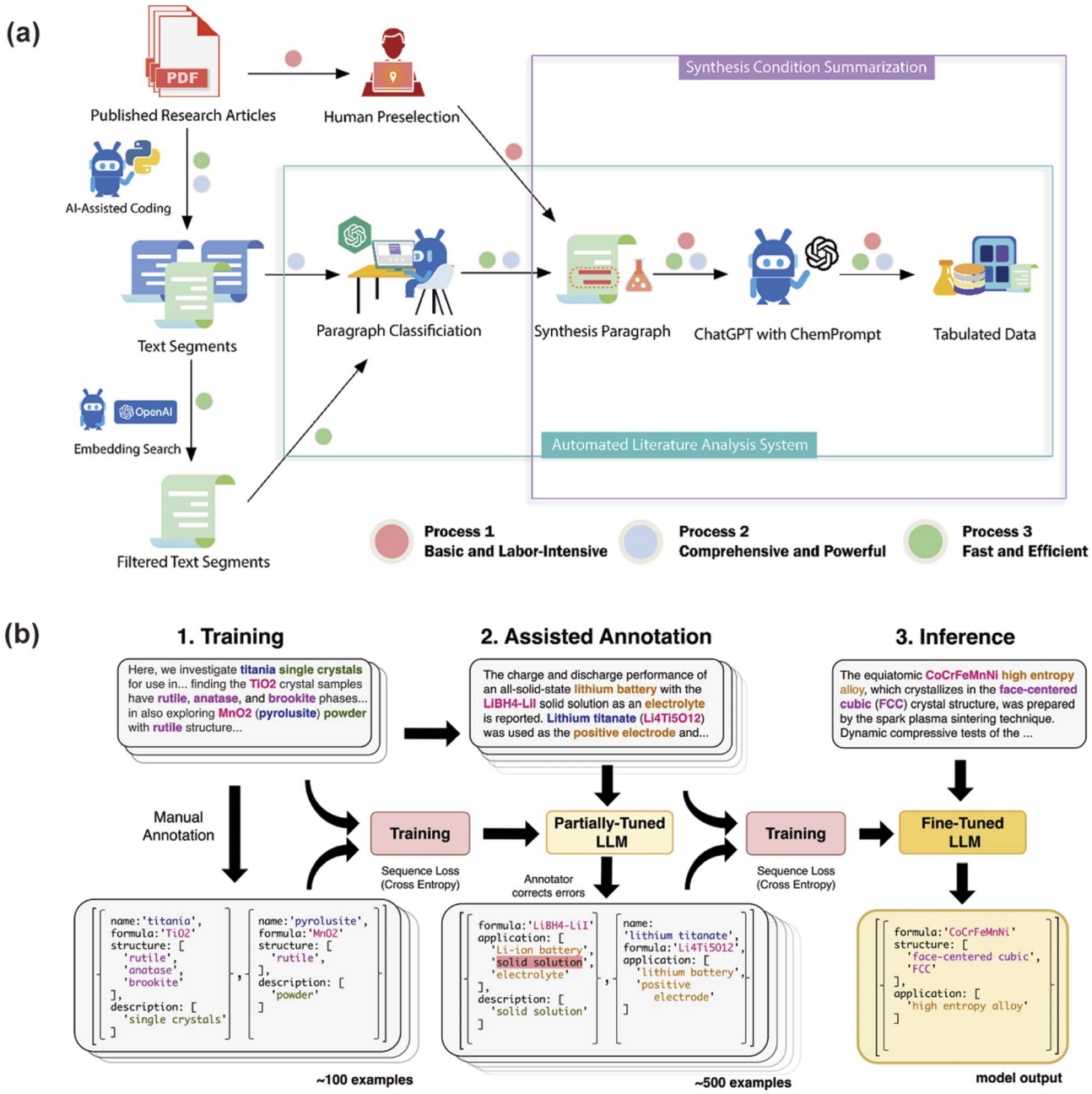

Fig. 3 Representative examples of LLM-driven literature mining systems for porous materials (a) a prompt-guided workflow that performs synthesis-condition extraction and tabular summarization from unstructured scientific text. Adapted from ref. 51. Copyright 2023 American Chemical Society. (b) A fine-tuned Named Entity Recognition and Relation Extraction (NERRE) model that produces structured, relational scientific records aligned with hierarchical schemas. Adapted from ref. 55 and licensed under CC-BY 4.0.

a reproducible workflow for literature-to-database conversion. The authors also examined hallucination behavior, identifying two primary error modes: fabrication of plausible synthesis conditions for non-existent MOFs through name-based pattern extrapolation, and incorrect factual associations such as misidentification of a MOF's metal center. To address these issues, the prompting framework was designed to reduce overconfident responses by allowing explicit abstention when information was uncertain and by constraining answers to a curated synthesis dataset derived from the literature. However, the extraction performance strongly relies on clearly structured or semi-structured experimental descriptions, and the method remains limited in handling loosely written narrative text or heterogeneous reporting styles without explicit human preselection.

Beyond prompt engineering, Zhang *et al.*[33] fine-tuned GPT-3.5 and open-source models (LLaMA-3, Mistral) for chemistry-specific text mining tasks, including compound and reagent recognition, reaction-role classification, and MOF-specific synthesis information extraction (Paragraph2MOFInfo).

Notably, strong performance was achieved using only a few hundred manually annotated paragraphs, with exact-match accuracy exceeding 80% for the MOF-specific extraction task. The authors acknowledge that LLMs may hallucinate by generating outputs inconsistent with established chemical knowledge, and show that supervised fine-tuning substantially reduces such unintended generations compared to prompt-only approaches. These results demonstrate that task-specific domain adaptation through fine-tuning markedly enhances consistency, reduces hallucination, and minimizes the need for extensive prompt engineering. Together, these advances establish the foundation for large-scale, autonomous chemical text mining and data-driven innovation grounded in experimental literature. Nevertheless, the scope of extractable information remains constrained by task-specific schemas and annotated training data, limiting direct transferability to unseen synthesis variables or broader classes of porous materials.

While textual information conveys much of the experimental context, crucial experimental data in porous-materials research often reside in figures such as adsorption isotherms, PXRD diffractograms, TGA curves, and microscopy images. Zheng *et al.*[52] demonstrated a vision-enabled large language model (GPT-4V) for multimodal data mining in reticular chemistry. Using natural-language prompts, GPT-4V analyzed 6240 images from 346 MOF papers and successfully classified and, to a substantial extent, interpreted key figure types including nitrogen isotherms, powder x-ray diffraction (PXRD) patterns, thermogravimetric analysis (TGA) curves, and structural diagrams. The model achieved classification accuracies above 94% ($F_1 \approx 93$–95%) for major categories and was able to infer contextual attributes such as gas type, temperature, and thermal stability trends. However, the reported errors were not entirely random. Misclassification occurred recurrently in visually ambiguous or broadly defined categories. In particular, IR and NMR spectra were sometimes incorrectly identified as gas sorption isotherms within the "other isotherm" category, reflecting structural similarity between line-plot formats. On pages containing multiple plot types, partial omissions were also observed, where one coexisting figure type (*e.g.*, TGA alongside PXRD) was missed. In addition, tasks requiring visual inference such as identifying hysteresis behavior or estimating saturation plateaus from adsorption curves showed lower accuracy than extraction of explicitly annotated textual values. Although the study did not systematically quantify hallucination instances, the authors explicitly incorporated prompt-level safeguards to minimize unsupported generation, instructing the model to rely strictly on information present in the page image and to return "N/A" when relevant data were absent or ambiguous. This design aimed to reduce hallucination-type errors, particularly fabrication of non-existent numerical or contextual information. This integration of visual and textual modalities represents a critical step toward fully digitalized experimental knowledge extraction, enabling autonomous retrieval of structure–property information from both narrative and graphical sources. Despite its strong performance, the accuracy of GPT-4V remains sensitive to prompt formulation, and misclassification was more frequently observed for visually ambiguous or broadly defined figure categories, while precise quantitative value extraction from plots remains challenging.

In some cases, literature-mined data have been further exploited to enable data-driven studies, rather than remaining as static databases. Kang *et al.*[53] developed L2M3 (Large Language Model MOF Miner), a large-scale autonomous pipeline designed to extract and standardize textual and tabular information from MOF literature. Processing over 40 000 publications, L2M3 integrates specialized agents for table parsing, synthesis-condition recognition, and property extraction under a central controller, yielding a structured database of 32 properties and 21 synthesis categories linked to experimental entries in the CSD. The pipeline achieved F_1 scores above 0.9 for extraction. The authors explicitly acknowledge the potential for hallucination and inconsistent extraction in LLM-based workflows and implement mitigation strategies including multi-stage agent chaining, structured JSON-constrained outputs, temperature control, and metadata cross-checking with the CSD to minimize fabricated or misassigned information. Leveraging this curated synthesis-condition dataset, the authors further developed a synthesis condition recommender system that suggests plausible reaction conditions based on given synthesis condition, demonstrating how literature-mined data can be transformed into an active, data-driven tool for guiding MOF synthesis. The fine-tuned recommender achieved a median recommendation score of ~0.83, significantly outperforming prompt-based and rule-based baselines. However, because extraction quality ultimately depends on the underlying LLMs, residual errors and inconsistencies may propagate across the multi-agent pipeline, particularly at large scale, necessitating continued verification and post-processing for high-confidence applications.

In addition, Hu *et al.*[54] demonstrated that LLM-derived knowledge can be converted into machine-readable datasets suitable for downstream analysis. In their study, the ChatGPT-4 API was used to extract and standardize synthesis parameters, pore characteristics, elemental composition and CO_2 uptake values for porous carbon materials from unstructured text, thereby establishing a structured experimental database of porous carbon adsorption data composed of over 10 000 individual entries. The resulting structured dataset was later used in an AutoML framework to explore synthesis–performance relationships. Although the optimization and design steps were conducted by conventional machine-learning models rather than the LLM itself, the study demonstrates how LLM-based extraction can act as the data-standardization layer enabling automated modeling, trend identification and hypothesis generation, representing a transitional stage toward fully autonomous research pipelines. At the same time, the authors note that LLM-extracted datasets may contain redundancy or missing entries, requiring prior evaluation and manual verification, and that experimental validation remains essential to confirm trends inferred from the automatically generated data.

Recently, LLM-assisted data mining has been extended to covalent organic frameworks (COFs) for synthesis-condition prediction. Zhao *et al.* constructed a COF synthesis database (SYN–COF) by using the large language model Deepseek-R1 to

extract monomer identities, reaction temperatures, times, and solvent systems from 609 literature sources, yielding 587 curated solvothermal entries.[55] Using prompt engineering, Deepseek-R1 achieved 97.19% extraction accuracy on manually evaluated samples, outperforming a BERT-CRF model by approximately 14% while requiring no annotated training data and demonstrating over 20-fold higher efficiency than manual extraction. The extracted data were encoded *via* SMILES-derived molecular fingerprints and used to train multiple ML models, with XGBoost achieving $R^2 = 0.88$ for temperature and $R^2 = 0.47$ for reaction-time prediction. To account for the multiplicity of viable synthesis regimes, additional classification models were constructed for discretized parameter ranges and common solvent combinations. The predicted conditions were experimentally validated through the successful synthesis of a previously unreported imine-linked COF (BPQD-TPDA) under model-recommended conditions (119 °C, 90 h, *o*-DCB/*n*-BuOH), whose crystallinity and microporosity were confirmed by PXRD and nitrogen sorption analyses. However, the framework is largely confined to frequently reported dual-monomer solvothermal systems, reflecting the limited availability and uneven distribution of reported COF synthesis data, particularly for novel linkage chemistries.

In addition to direct value extraction, several studies have explored the extraction of relational and structured knowledge from materials literature. Dagdelen *et al.*[56] developed an end-to-end joint named entity recognition and relation extraction (NERRE) framework by fine-tuning GPT- and LLaMA-based architectures to automatically capture hierarchical relationships among entities in materials literature. The assisted annotation and fine-tuning workflow for joint entity and relation extraction is schematically illustrated in Fig. 3b. One benchmark task focused on MOFs, where models were trained on several hundred abstracts to identify MOF names, chemical formulae, guest species, applications, and descriptive attributes, and to organize these entities into a predefined JSON-based hierarchical schema. This approach effectively generated structured records such as MOF-guest-application triads, providing a relational view of MOF chemistry and function that is difficult to achieve with rule-based or BERT-like pipelines. The fine-tuned LLMs demonstrated strong performance (*e.g.*, $F_1 \approx 0.57$ for name–application and 0.62 for name–guest relations) and showed the ability to normalize and correct chemical entities automatically. The authors explicitly acknowledge hallucination as a limitation, noting that the model may generate or infer chemical names or formulae not explicitly present in the input text. Although such inferences can be chemically plausible, they are considered inappropriate for strict information extraction, as extracted entities should be directly grounded in the source passage. For porous materials research, such structured and relation-aware representations provide a foundation for building knowledge graphs linking literature-derived insights, thereby supporting large-scale data integration, semantic search, and data-driven hypothesis generation. Nonetheless, the method occasionally produces formatting inconsistencies and hallucinated relations not explicitly supported by the source text, indicating the continued need for schema verification and human oversight in high-stakes applications.

Rather than focusing solely on relation extraction, LLMs have been applied to more complex reasoning tasks that require evidence aggregation and verification. Ansari and Moosavi[57] proposed Eunomia, an autonomous chemistry agent that advances LLM-based extraction from static text parsing to dynamic reasoning and verification. Eunomia uses a large language model (LLM) in a ReAct-style agent setup, allowing the model to think step by step and decide when to search documents or verify its own answers. This design makes it possible to handle multi-step information extraction and to reason over entire papers without fine-tuning. Importantly, the authors explicitly discuss hallucination as a key limitation of LLM-based systems, defining it as the generation of unsupported or fabricated information, and incorporate a Chain-of-Verification (CoV) module to re-examine extracted evidence before producing final outputs, thereby reducing ungrounded or incorrect content. When benchmarked on three information-extraction tasks of increasing complexity (solid-state doping relations, MOF chemical-formula and guest-species identification, and MOF water-stability classification), Eunomia achieved zero-shot performance comparable to or exceeding fine-tuned LLMs. For MOF formula extraction, Eunomia increased the F_1 score from 0.424 (fine-tuned baseline) to 0.606, while showing high recall (0.923) in guest-species identification. In the most challenging task of MOF water-stability classification at the full-paper level, Eunomia achieved a ternary accuracy of 0.91 with an information recovery yield of 86.2%. This work highlights a broader shift from fine-tuned or prompt-engineered models toward tool-augmented, self-verifying agents for materials information extraction, demonstrating the potential for more accurate and scalable database generation from the literature. Despite these advantages, Eunomia's performance remains highly dependent on clear task decomposition and prompt design, and the added system complexity introduced by multi-step reasoning and tool usage may affect robustness in the absence of carefully engineered guidance.

2.2.2 Applications of LLMs in porous materials: beyond data extraction. Building upon the success of LLMs in literature-based data mining, recent research has expanded their applications toward knowledge reasoning, property prediction, and generative design in porous materials such as MOFs and zeolites. These works collectively demonstrate that LLMs can function not only as information extractors but also as knowledge interpreters and creators. In porous materials workflows, LLMs are best viewed as complementary components that integrate unstructured knowledge with established predictive and optimization models. While most applications emphasize reasoning and coordination, fine-tuned models have also demonstrated the capacity for direct structure–property prediction from symbolic representations. Their functional role therefore depends on task design rather than replacing existing methods.

To assess this expanded functional scope, several studies have focused on systematically evaluating the capabilities of LLMs across a broad range of MOF-related research tasks. Bai

et al.[58] systematically evaluated six open-source LLMs, including LLaMA2-7B, ChatGLM2-6B, and Falcon-7B, across a comprehensive suite of MOF-related tasks such as chemistry knowledge, MOF database reading, experiment design, computational script generation, data analysis, and property prediction (Fig. 4a). Their results showed that moderate-sized models (6–7 billion parameters) demonstrated reasonable understanding of domain-specific concepts and could generate usable experimental designs and simulation inputs with performance comparable to GPT-3.5 in several qualitative and semi-structured tasks, including MOF knowledge recall, database querying, and the generation of experimental designs and computational scripts. Among the evaluated models, LLaMA2-7B and ChatGLM2-6B consistently exhibited the most balanced performance across these tasks, combining reliable domain understanding with moderate computational requirements. Despite generally strong performance in knowledge retrieval and research-assistance tasks, the evaluated models exhibited limited MOF-specific depth, showed constrained reliability in property-related reasoning, and often generated experimentally plausible but insufficiently specific suggestions without domain-specific fine-tuning. This study provided a systematic comparison of multiple open-source models and offered practical guidance for improving their fine-tuning and domain adaptation in future porous-materials research.

To support deeper reasoning and knowledge-grounded workflows, recent efforts have focused on constructing large-scale question–answer corpora that encapsulate reticular-chemistry knowledge in a machine-interpretable form. Rampal *et al.*[59] introduced RetChemQA, a large-scale benchmark dataset designed to evaluate the reasoning and comprehension abilities of LLMs in reticular chemistry. RetChemQA comprises around 90 000 single- and multi-hop question–answer pairs generated from approximately 2500 MOF-related publications using GPT-4-Turbo. The dataset spans factual, reasoning, and true/false question types, enabling fine-grained assessment of model understanding across scientific tasks. In evaluating model reliability, the authors explicitly address hallucination by defining it as the generation of out-of-context Q & A pairs and introduce quantitative metrics, including hallucination rate and hallucination capture rate, to systematically evaluate and analyze such behavior. Moreover, it provides a foundation for developing automated prompt optimization frameworks such as DSPy,[100] facilitating iterative improvement of LLM performance without manual intervention. This work establishes a shared benchmark for evaluating complex reasoning in reticular chemistry. Although RetChemQA provides a large-scale, reasoning-oriented QA corpus, the generated question–answer pairs and downstream model outputs still require human validation, particularly for synthesis feasibility and structural correctness.

Beyond enabling knowledge representation and reasoning workflows, an important question is whether such language-based learning can be extended to structure–dependent property prediction in porous materials. Wu and Jiang[60] presented one of the first demonstrations of applying a fine-tuned general-purpose LLM (Gemini-1.5) to predict the hydrophobicity of MOFs. In their framework, MOF structures were represented as chemical strings (SMILES and SELFIES, Self-Referencing Embedded Strings) and used to fine-tune Gemini-1.5 as a supervised end-to-end classifier, enabling the model to learn latent chemical language patterns of structural motifs. Importantly, the fine-tuned Gemini directly outputs hydrophobicity class labels from symbolic MOF representations, without relying on external feature engineering or downstream machine-learning predictors. The fine-tuned Gemini achieved a weighted accuracy of up to 0.78 for binary classification and 0.73 for quaternary classification, outperforming descriptor-based SVM models built on pore and RAC features, which reported weighted accuracies on the order of 0.75 (binary) and 0.70 (quaternary). The model further retained robust performance even under moiety-masking (partial-input) conditions. This study highlights that with minimal domain-specific retraining, LLMs can infer physicochemical properties directly from symbolic chemical representations, bridging the gap between text-based learning and quantitative materials prediction. However, the model exhibited reduced predictive performance when applied to solvent- or ion-containing MOFs outside the training distribution, highlighting challenges in out-of-distribution generalization.

In addition to predicting material properties, LLMs can also actively contribute to the design and synthesis of entirely new materials. Zheng *et al.*[61] introduced an LLM-based generative design framework for MOF linker mutation, coupling data curation, model fine-tuning, and experimental synthesis for water-harvesting (Fig. 4b). Using a curated dataset of 3943 linker-editing examples covering four mutation categories, including substitution, insertion, replacement, and positioning, the fine-tuned model achieved significantly higher accuracy (84.8%) and recall (93.9%) in generating valid chemical structures compared with the base GPT-3.5 and GPT-4 models. The authors further note that base models sometimes produced hallucinated SMILES strings that were syntactically plausible but chemically invalid or inconsistent with the specified mutation instructions, highlighting the need for task-specific fine-tuning. The model proposed new linker variants predicted to enhance water-harvesting performance, which were subsequently synthesized into the LAMOF series (LAMOF-1 to LAMOF-10). These MOFs feature heteroatom-substituted linkers and demonstrate record water uptake (up to 0.64 g g^{-1}) with tunable humidity response (13–53% RH). This study provided a practical demonstration of LLM-driven reticular chemistry, where fine-tuned language models can act as AI co-designers that accelerate the generation and synthesis of functionally enhanced, synthetically feasible MOFs. However, the model can generate new structures only within the space defined by combinations of linker-editing rules represented in the training data.

2.3 Autonomous systems: from human-in-the-loop to self-driving labs

The advent of LLMs has introduced an innovative paradigm not only for data extraction but also for data analysis across diverse

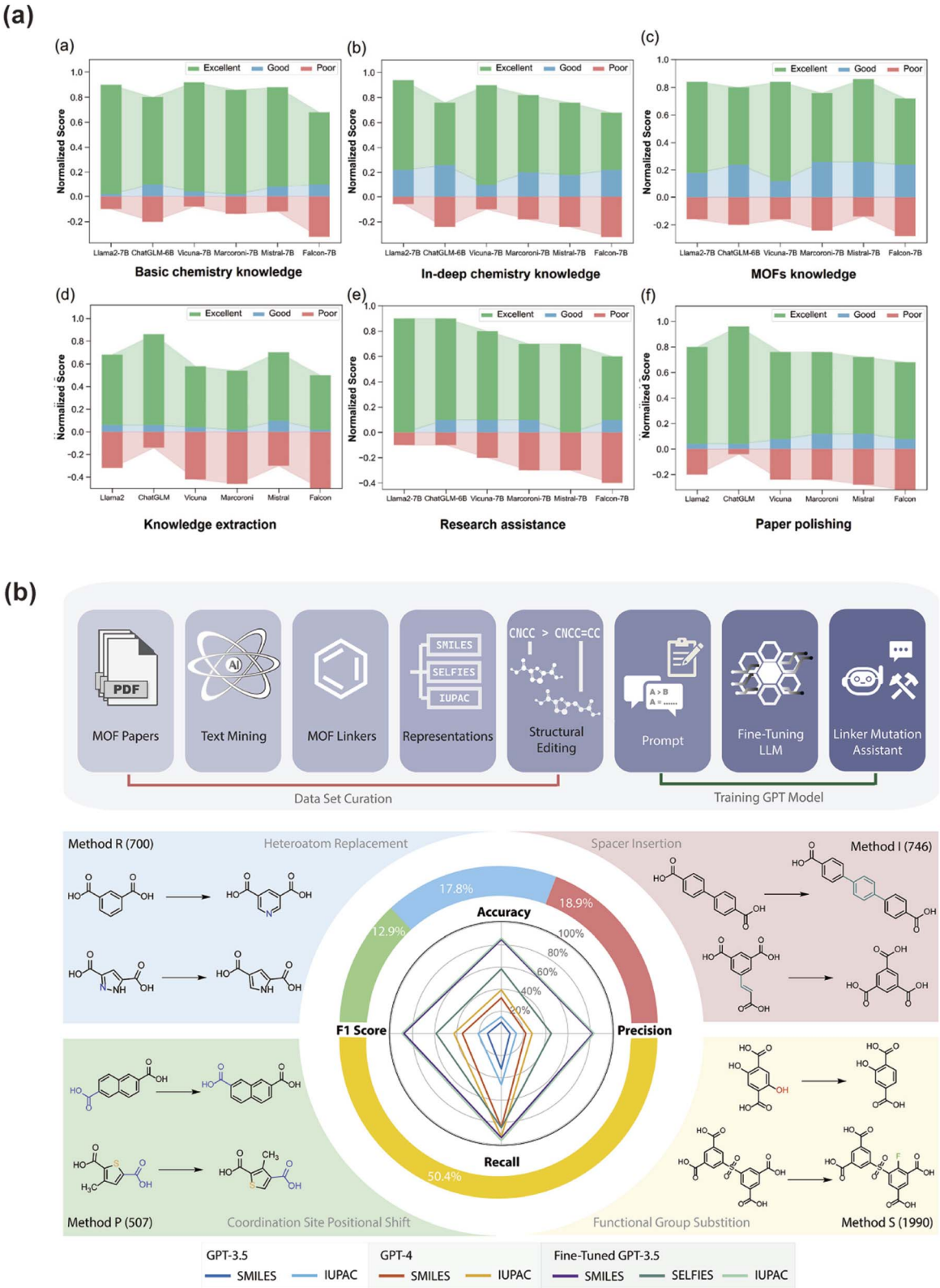

Fig. 4 Expansion of LLM applications in porous materials. (a) Evaluation of open-source LLM capabilities for chemistry- and MOF-focused tasks. Adapted from ref. 57. Copyright 2024 American Chemical Society (b) LLM-assisted MOF linker mutation workflow with experimental validation. Adapted from ref. 60. Copyright 2023 American Chemical Society.

materials research fields.[101] By enabling closed-loop automation that integrates data extraction, hypothesis generation, experimental execution, and iterative feedback, these models support increasingly autonomous workflows for materials design, synthesis, and optimization with minimal human intervention.[102] Within this framework, LLMs act as the cognitive core, coordinating data, models, and robotic systems to predict synthetic routes, generate executable workflows, and guide adaptive decision-making.[103] As feedback and control become partially autonomous, this integration culminates in the self-driving laboratory (SDL), where reasoning, execution, and evaluation operate in a seamless discovery cycle that accelerates materials innovation.[104]

These LLM-driven automation strategies are being actively and broadly adopted across diverse materials and chemistry domains, suggesting broad applicability and the potential for transformative impact. In the field of quantum chemistry, significant progress has been made in democratizing access to sophisticated computational tools. Gadde *et al.* introduced AutoSolvateWeb,[105] a chatbot-assisted platform that employs Google Dialogflow CX[106] to guide non-expert users through multistep quantum mechanical/molecular mechanical (QM/MM) simulations of explicitly solvated molecules, while operating on cloud infrastructure to completely eliminate hardware configuration barriers.[105] Zou *et al.* developed El Agente Q, an LLM-based multi-agent system which implements a hierarchical multi-agent architecture where specialized agents collaboratively handle dynamic task decomposition, adaptive tool selection, and post-analysis in quantum chemistry.[107] El Agente Q reported an 87% task success rate in university-level quantum chemistry benchmarks, thereby enabling users to execute complex workflows from natural language prompts without external intervention.

In experimental synthesis, the integration of LLMs with robotic laboratory systems has demonstrated increasing levels of autonomous operation. Song *et al.* introduced ChemAgents, a hierarchical multiagent-driven robotic AI chemist.[108] It is powered by an LLM (Llama-3.1-70B) that coordinates four specialized agents: the Literature Reader, Experiment Designer, Computation Performer, and Robot Operator. A key achievement was the discovery and optimization of high-performance metal–organic high-entropy catalysts (MO-HECs) for the oxygen evolution reaction (OER). Similarly, Huang *et al.* reported a natural-language-interfaced robotic platform for inorganic materials translating synthetic procedures directly into executable operations.[109] This platform autonomously synthesized 13 compounds across four material classes: coordination complexes, MOFs, nanoparticles, and polyoxometalates. Furthermore, through AI copilot-assisted exploration, the system discovered four previously unreported Mn–W polyoxometalate clusters (specifically Mn_4W_{18}, Mn_4W_8, Mn_8W_{26}, and $Mn_{57}W_{42}$). These advances highlight the potential of human-AI collaboration in accelerating materials discovery. Together, these studies demonstrate how LLM-driven automation bridges reasoning and execution, advancing chemistry toward fully autonomous discovery.

Building upon these developments, we now consider how similar approaches are emerging in porous materials research. In the following section, we examine representative examples of human-in-the-loop, closed-loop automation and LLM–robotics integration.

2.3.1 LLM as research co-pilots. In porous materials research, LLMs are increasingly adopted as research co-pilots that support discovery across both experimental and computational workflows. Across the literature, these systems are commonly used to assist with literature mining, knowledge extraction, and the proposal of synthesis, optimization, or design strategies, while researchers remain involved in evaluating feasibility and interpreting results. At the same time, many studies integrate LLMs more tightly into computational pipelines, where models iteratively generate candidates, propose design modifications, and evaluate outcomes using simulations, surrogate models, or automated screening criteria. In such settings, LLMs play an active role in guiding the exploration process within the computational loop, while human involvement is often limited to stages outside the automated cycle, such as defining initial constraints or interpreting generated results.

One prominent example is the GPT-4 Reticular Chemist, a pioneering framework designed to guide the discovery of new MOFs through seamless, conversational collaboration between a chemist and an LLM.[62] Operating entirely through natural language, it eliminates the need for specialized coding skills, making advanced LLM-based reasoning accessible to any researcher in the field. The system's workflow is structured into three distinct but interconnected phases: (1) reticular ChemScope, which generates a high-level conceptual research plan based on user input; (2) reticular ChemNavigator, which evaluates experimental outcomes and proposes the next set of actions with supporting rationale; and (3) reticular ChemExecutor, which produces detailed step-by-step laboratory protocols for execution. A feedback mechanism links these stages. After each experiment, the researcher provides a natural-language summary of the outcome, which the LLM incorporates as contextual memory to refine subsequent decisions. To mitigate potential relation-type hallucination, where the LLM may infer incorrect relationships or provide flawed interpretations of experimental data such as NMR or TGA, the authors adopted a human-led analysis strategy supplemented by conventional ML algorithms to ensure analytical reliability. The workflow was applied to the exploration of the MOF-521 isoreticular series, guiding progression from linker synthesis and reaction-condition refinement to final characterization, indicating how LLM-assisted planning can complement human experimental execution. The authors also report good agreement between simulated and experimental PXRD patterns, with experimentally measured surface areas consistent with computational predictions. Furthermore, the authors discuss reproducibility at the level of the discovery process rather than in terms of identical textual outputs. In their study, the total number of prompt iterations required to reach optimized synthesis and characterization stages was similar across different MOF-521 derivatives, despite variations in

intermediate decision paths. This observation suggests that, in conversational LLM-driven workflows, reproducibility has been discussed in terms of achieving similar overall iteration counts and workload, even when intermediate decision paths differ.

Beyond the GPT-4 reticular Chemist, another representative example of HITL reasoning is the MOFsyn agent, an advanced framework that enhances the synthesis optimization of MOF catalysts through a retrieval-augmented generation (RAG) framework (Fig. 5a).[63] MOFsyn integrates three synergistic components: (1) the Data Automatic Analyzer, which autonomously performs ML-based analysis of catalytic data without coding expertise; (2) the Material Mechanism Analyzer, which combines RAG-based literature querying with real-time online searches to provide mechanistic insights and synthesis recommendations; and (3) the Experimental Protocol Navigator, which guides iterative human–AI collaboration through adaptive experimental design. Underlying the interaction among these components, MOFsyn employs chain-of-thought (CoT) reasoning to decompose complex synthesis optimization tasks into explicit, sequential reasoning steps.

The RAG framework in MOFsyn grounds synthesis reasoning in domain-specific knowledge curated from the literature and supplemented by real-time web retrieval, addressing the limitations of static LLM training. The local corpus was constructed

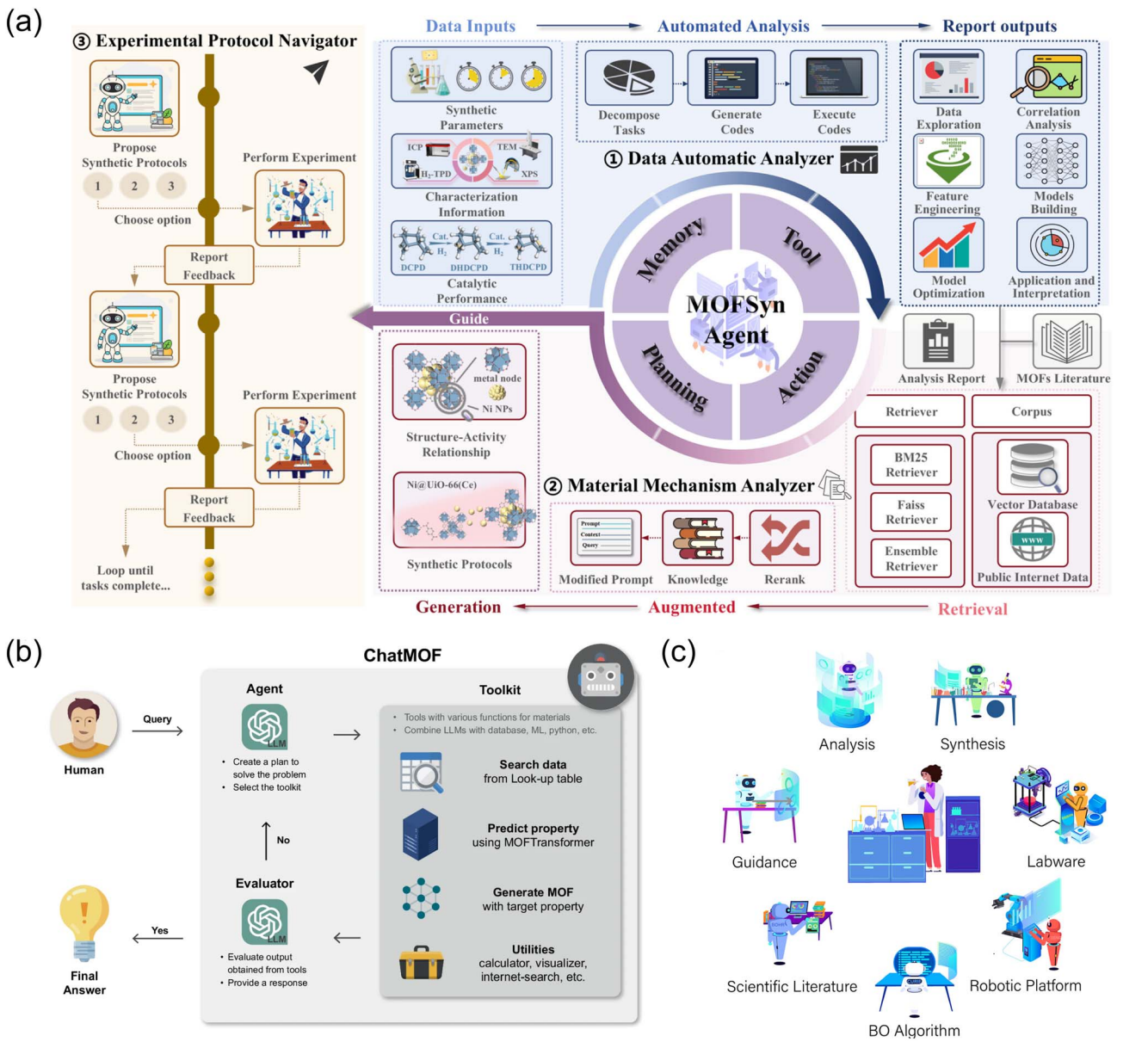

Fig. 5 Representative examples of large language model (LLM)–enabled workflows discussed in Section 2.3. (a) MOFsyn agent illustrating an LLM-enabled human-in-the-loop framework for synthesis guidance. Adapted from ref. 62. Copyright 2025 American Chemical Society (b) autonomous MOF design framework (ChatMOF) combining LLM reasoning, property prediction, and inverse design. Adapted from ref. 23 and licensed under CC-BY 4.0. (c) An example of robotics-assisted experimentation: the ChatGPT Research Group multi-agent workflow. Adapted from ref. 24 and licensed under CC-BY 4.0.

from 508 records retrieved from the Web of Science Core Collection on June 17, 2024, using targeted search queries focusing on Ce-based MOFs for hydrogenation, Ni nanoparticle-catalyzed hydrogenation, structure–activity relationships of Ce-MOF-supported Ni catalysts, and DCPD hydrogenation mechanisms. Retrieval over the curated corpus employs an ensemble strategy that combines Facebook AI similarity search (FAISS, for dense embedding retrieval)[110] and Best Matching 25 (BM25, for sparse lexical retrieval)[111] with equal weighting (0.5/0.5). This design leverages the complementary strengths of sparse retrieval in precisely matching specific chemical formulas and identifiers and dense retrieval in capturing broader semantic contexts and thematic analogies, effectively balancing the trade-off between keyword-level accuracy and contextual relevance. The ensemble retriever[112] first identifies the top-10 candidate documents from the local corpus. These are then combined with web-retrieved candidates to form a preliminary pool. Retrieval precision is enhanced through a two-stage reranking pipeline: cosine similarity filtering (threshold = 0.8) followed by cross-encoder semantic reranking using Cohere's rerank-english-v2.0 model. The multi-stage design mitigates the inclusion of weakly relevant documents from initial retrieval, improving contextual precision before prompt injection. After reranking, the five most relevant documents are selected and injected into the final prompt as contextual grounding.

To optimize LLM prompting for RAG reasoning, MOFsyn employs a specialized prompt-engineering framework called R.O.S.E.S (role definition, objective statement, scenario description, expected solution, and structured output). In both the Retrieval-Augmented Generation Assessment (RAGAS)[113] and a 100-question MOF materials test, GPT-4o achieved the highest overall RAG performance than Deepseek-V3, GLM-4-Flash-250414, and Qwen2.5-MAX under the evaluated configuration (Fig. 6a). Retrieval quality was quantified using standard RAGAS metrics, faithfulness, answer relevance, context recall, and answer similarity, alongside performance on the domain-specific test set. For GPT-4o, ensemble retrieval improved performance from 89% (no retriever) to 95%. In practical deployment, online retrieval typically responds within 1–2 s. However, multi-stage retrieval and reranking introduce additional latency and token overhead, particularly as corpus size increases, which should be explicitly considered in reproducibility and scalability evaluations. Regarding reproducibility, the authors explicitly controlled generative variability by setting the inference temperature to 0.1, with the stated goal of reducing stochastic variability and obtaining accurate and consistent outputs during synthesis reasoning.

In practical application, the system optimized the synthesis of a Ni@UiO-66(Ce) catalyst. MOFsyn identified limitations in conventional one-pot reduction routes, and recommended a two-step low-temperature reduction pathway, which later yielded improved selectivity and conversion. These improvements were further contextualized by confirming structural integrity of UiO-66(Ce) *via* XRD and validating the Ni^0 active species, with the effects of excess $NaBH^4$ discussed in accordance with literature-reported mechanisms. Taken together, this case illustrates how retrieval-grounded reasoning can support iterative decision-making within a human–AI synthesis workflow.

A representative example of LLM-integrated computational exploration is found in LLM-guided design of organic structure-directing agents (OSDAs) for zeolite synthesis.[64] Although not directly targeting porous frameworks themselves, OSDAs play a pivotal role in zeolite synthesis, which represents one of the most important classes of porous materials.[114] OSDAs are quaternary ammonium cations that guide zeolite crystallization and stabilize specific framework topologies.[115] In this work, Ito *et al.*[64] developed an iterative closed-loop workflow that couples GPT-4 with atomistic simulations to progressively refine OSDA candidates. The LLM generates new OSDA molecules starting from the simple prototype tetraethylammonium (TMA). Empirical domain filters, such as a carbon-to-nitrogen ratio between 4 and 20 and fewer than five rotatable bonds, are applied, and unsuitable molecules are rejected with natural language feedback returned to the LLM. Screened molecules undergo stabilization-energy calculations to obtain zeolite affinity scores, and qualified candidates are stored in a growing database. The authors further examined how output variability depends on the model temperature by systematically analyzing GPT-4-generated OSDA molecules under fixed input prompts. At a temperature of 0.3, the model produced relatively deterministic outputs, but a noticeable fraction of the generated molecules were duplicated or closely resembled the inputs, leading to limited diversity. At temperatures below 0.3, the fraction of unique molecules further decreased, indicating inefficient exploration of chemical space. By contrast, increasing the temperature to intermediate values (0.7–1.1) resulted in more diverse molecular proposals while maintaining a high fraction of syntactically valid SMILES. At higher temperatures exceeding 1.3, the fraction of parsable SMILES decreased markedly due to fabrication-type hallucination, where the model produced nonsensical or near-random character sequences leading to syntactically invalid SMILES strings. Based on these observations, the authors adopted a stochastic sampling strategy by randomly selecting temperatures from 0.7 to 1.1 during iterative design, explicitly balancing reproducibility, molecular diversity, and exploration efficiency within the closed-loop workflow.

Using this temperature-controlled design strategy, the workflow was tested on three cage-type zeolites: CHA,[116] AEI,[117] and ITE.[118] The resulting candidates showed stabilization energies comparable to or higher than those reported for experimentally validated OSDAs. However, synthesizability was not considered in the current implementation, and the model occasionally failed to infer realistic synthetic pathways for structurally complex candidates. These observations indicate that further integration of synthesis-oriented considerations would be beneficial for translating computationally proposed OSDAs into experimentally realizable systems.

Beyond tightly coupled simulation feedback loops, related studies emphasize LLM-driven orchestration across complex computational workflows. In this direction, Ding *et al.* proposed SciToolAgent, an LLM-based framework designed to autonomously orchestrate hundreds of specialized scientific tools across biology, chemistry, and materials science through the

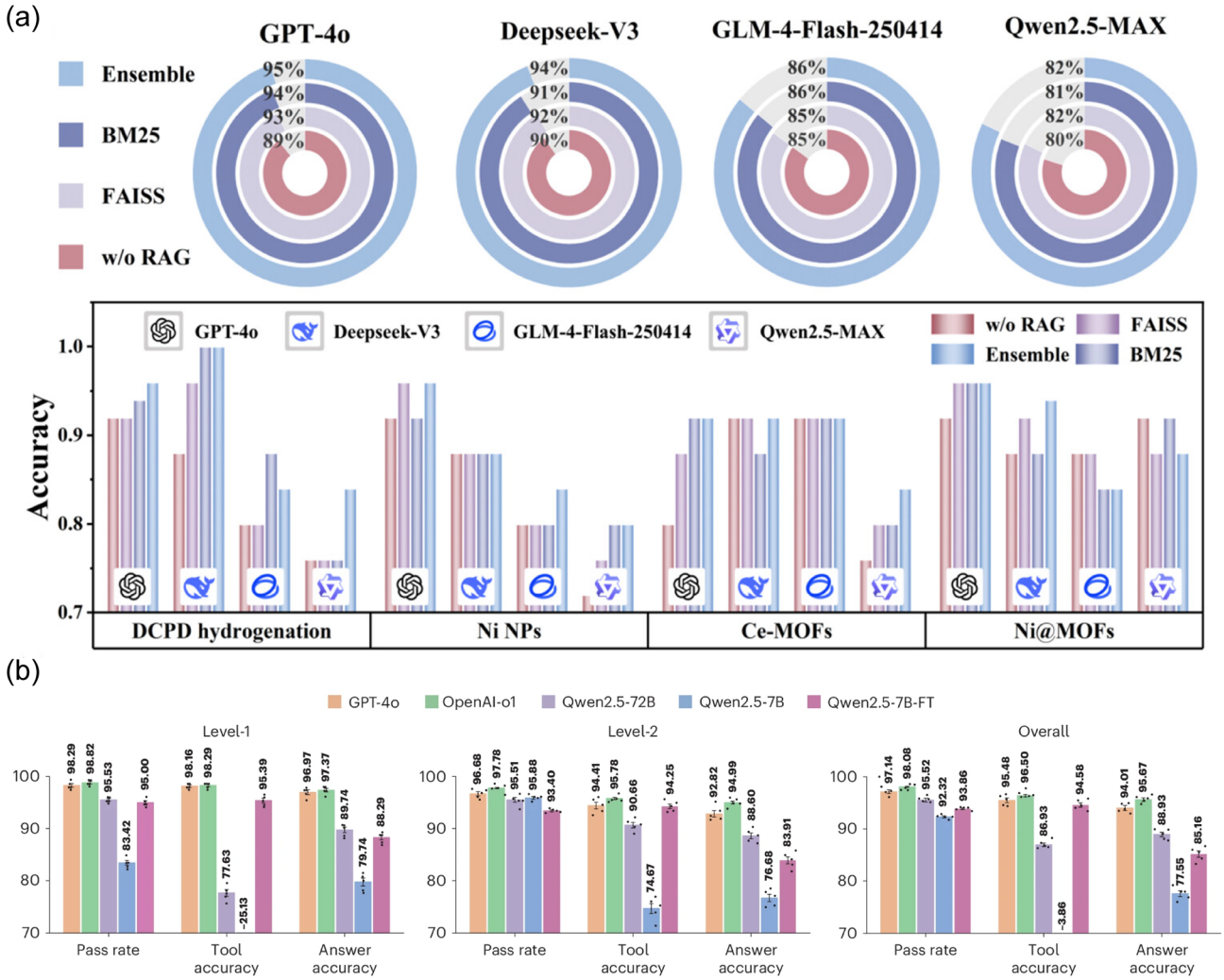

Fig. 6 Benchmarking the performance of different large language models (LLMs) used in emerging LLM-enabled frameworks for porous materials research. (a) Impact of retrieval augmentation in MOFsyn, where ensemble retrieval provides the highest accuracy overall (top), with model- and category-dependent variation observed across MOFs materials exam (bottom). Adapted from ref. 62. Copyright 2025 American Chemical Society (b) foundation model benchmarking in SciToolAgent shows OpenAI-o1 as the best-performing model, with GPT-4o providing the most efficient accuracy–cost balance. Performance metrics represent mean values over repeated benchmark evaluations ($n = 5$), with GPT-4o showing strong overall performance across task categories. Adapted from ref. 64. Copyright 2025 Springer Nature America.

scientific tool knowledge graph (SciToolKG).[65] The core of SciToolAgent lies in leveraging SciToolKG to mediate communication between the LLM and integrated scientific tools, enabling intelligent tool selection and execution through graph-based RAG. The system consists of three main modules: the planner, executor, and summarizer. When the summarizer identifies suboptimal outcomes, it prompts the planner to refine the tool chain, reducing trial-and-error costs and iteratively optimizing until satisfactory results are achieved.

SciToolAgent reported 94% accuracy on benchmark datasets, outperforming previous models by approximately 10%. Among various foundation models tested, OpenAI o1 showed the highest absolute accuracy, whereas GPT-4o offered the most favorable balance between performance and computational cost and was therefore selected as the default model (Fig. 6b). Consistent performance trends observed across repeated benchmark evaluations ($n = 5$) and multiple foundation models provide an indirect indication of system-level reliability. The framework demonstrated versatility across four case studies: protein design, chemical reactivity prediction, synthesis planning, and MOF screening. In the MOF screening task, SciToolAgent autonomously identified a MOF with high thermal stability (above 400 °C), high CO_2 adsorption capacity (above 100 mg g^{-1}), and low synthesis cost (below ¥100). It then constructed and executed a sequential workflow integrating a machine-learning-based predictor (MOFSimplify[89]) and a molecular simulation tool (RASPA2 (ref. 119)) to select the optimal candidate. However, despite its promising performance, the system still faces scalability limitations due to the manual construction of SciToolKG and a reliance on proprietary models such as GPT-4o.

Recent approaches have pushed autonomous systems toward knowledge-grounded inverse design, where LLMs actively reason about how to modify material structures to achieve target properties. In this context, Ansari *et al.* introduced the dZiner framework, which exemplifies inverse design automation that explicitly incorporates domain knowledge into the generative loop.[66] It performs rational inverse design through a RAG loop that iteratively retrieves domain knowledge from literature, proposes chemical modifications, and assesses feasibility through LLM-based reasoning. Underlying this process, dZiner employs chain-of-thought (CoT) reasoning to translate retrieved design principles into explicit, stepwise modification strategies, preserving interpretable reasoning traces across successive iterations.

The workflow consists of three core capabilities: extracting design rules from text, generating new candidates in natural language, and evaluating them using surrogate models such as MOFormer[120] for CO_2 adsorption prediction. Powered by Claude 3.5 Sonnet and GPT-4o, dZiner supports both fully automated reasoning cycles and an optional human-in-the-loop mode, enabling collaboration when expert oversight is desired. When tasked with designing a linker for a high-uptake MOF, the model autonomously suggested nitrogen-rich and fluorinated heterocycles that enhanced the predicted CO_2 adsorption capacity, indicating that textual chemical knowledge can be coupled with surrogate model feedback in an automated loop. Here, the agent was implemented using a fixed inference temperature (0.3) and a ReAct-based chain-of-thought architecture, while candidate evaluation relied on surrogate models such as MOFormer.[120] To mitigate fabrication-type hallucination, where the LLM produced syntactically invalid SMILES corresponding to non-existent or chemically infeasible structures, the authors introduced an additional RDKit-based verification step to filter out invalid candidates before downstream evaluation.

While dZiner represents a major step toward automatic inverse design, its SMILES-based representation inherently oversimplifies complex frameworks such as MOFs, and its lack of multimodal understanding (*e.g.*, interpretation of images, schemes, or structural diagrams) remains a key limitation.

At a higher level of semantic abstraction, LLM-mediated autonomous systems aim to directly mediate between natural language descriptions of porous materials and structured representations, enabling language-driven reasoning and modular tool orchestration beyond manually implemented optimization pipelines. Building on recent advances in large-language-model-based materials research, Kang *et al.* developed ChatMOF, an autonomous AI system for predicting and generating MOFs (Fig. 5b).[23] ChatMOF couples LLMs (GPT-4 and GPT-3.5) with external materials databases, including CoRE MOF,[88] QMOF,[121] and DigiMOF,[83] as well as machine-learning-based property predictors such as MOFTransformer,[30] within a LangChain-based tool-orchestration framework inspired by the ReAct[122] and modular reasoning, knowledge and language (MRKL)[123] architecture.

Within this architecture, the ReAct paradigm enables ChatMOF to explicitly alternate between reasoning steps, where the LLM interprets user intent and plans subsequent actions, and action steps, in which specialized tools such as databases, predictors, or generative algorithms are invoked. Complementarily, the MRKL design allows the LLM to route queries to appropriate expert modules rather than attempting to solve all tasks internally, thereby leveraging domain-specific tools while maintaining a unified language-driven interface. During the reported evaluations, ChatMOF was operated without model fine-tuning, with a low temperature setting (0.1) used during inference.

Through this modular reasoning–acting pipeline, ChatMOF performs a range of natural-language-driven tasks, including data retrieval, property prediction, and inverse structure generation using genetic algorithms. Under GPT-4 evaluation, the system achieved up to 96.9% accuracy in search tasks, 95.7% in property prediction, and 87.5% in generative design benchmarks, excluding cases involving token-length overflow. These benchmark tasks primarily evaluate the correctness of tool selection and multi-step orchestration on curated query sets rather than end-to-end experimental research workflows. In this context, generative validity refers to producing structurally consistent MOF representations suitable for downstream computation and does not in itself imply synthetic accessibility or experimental performance improvement. By embedding LLMs as autonomous agents capable of coordinating databases, predictive models, and generative routines, ChatMOF exemplifies an emerging class of LLM-orchestrated discovery systems that bridge human dialogue, materials knowledge bases, and computational evaluation tools. Nonetheless, its performance remains constrained by token-length limitations, occasional reasoning failures, and limited generative diversity during MOF generation.

Together, these studies illustrate how LLMs function as research co-pilots across a range of computationally structured discovery workflows, from interactive synthesis planning to closed-loop exploration and workflow orchestration.

2.3.2 Toward Self-Driving Laboratories (SDL): integration of LLMs with robotics. While the preceding discussions examined human-in-the-loop frameworks and autonomous optimization pipelines, fully autonomous experimental exploration in porous materials research remains challenging due to the need for reliable synthesis, characterization, and safety control without continuous human oversight. Robotic systems offer a complementary route to extend automation toward experimental exploration, particularly by automating experimental exploration in porous materials research.

In recent years, robotic platforms have been combined with data-driven algorithms such as Bayesian optimization and genetic algorithms. These approaches have been used to efficiently search synthesis spaces and identify optimal conditions. For instance, Bayesian optimization accelerated the synthesis of ZIF-67,[124] and genetic algorithms guided the growth of HKUST-1 thin films of surface-anchored MOFs (SURMOFs) with high crystallinity and uniform orientation.[125]

More recently, a growing research direction has explored the integration of LLMs into robotic workflows, particularly at the level of experimental planning and strategy formulation. LLMs

can interpret scientific literature, propose synthesis strategies, and assist in automating experiment design. Although such systems do not yet operate in real time or without human input, recent examples such as ChatGPT Research Group,[24] MOFGen[67] and the green synthesis of Zn-HKUST-1 (ref. 68) illustrate how LLM-guided workflows can bridge conceptual reasoning and physical experimentation. These integrated pipelines mark the early stages of SDL systems, where AI agents increasingly support experiment design, execution, and evaluation as part of a unified workflow.

A representative example of this new paradigm is the ChatGPT Research Group[24] (Fig. 5c). Composed of seven AI agents integrated with Bayesian Optimization (BO), the system was designed to automate experimental planning and optimization for microwave synthesis. Each agent was assigned a specific scientific role through prompt engineering, including strategy planning, literature search, coding, robotic operation, labware design, safety inspection, and data analysis. In particular, the safety-related responsibilities were implemented through a dedicated chemistry consultant agent that provided guidance on laboratory safety precautions, such as handling microwave irradiation, pressure buildup, and chemical hazards during synthesis. Outputs were passed sequentially between agents, forming a multi-step reasoning pipeline rather than a single monolithic model.

The framework was applied to optimize synthesis conditions for MOF-321, MOF-322, and COF-323. For MOF-321, it identified optimal conditions after 120 robotic experiments conducted over 4.5 days. To evaluate synthesis outcomes, the authors introduced a crystallinity index (CI) defined as the ratio between the height of the primary diffraction peak and its full width at half maximum, with higher values corresponding to sharper and more crystalline products. The index increased steadily across iterations (1 to 36), indicating progressive convergence rather than random parameter exploration. Similar optimization trends were observed for MOF-322 and COF-323, suggesting that the workflow may be extensible to other porous materials systems. Despite its success, the framework remains semi-automated. The authors note that future integration with more advanced robotic platforms could further enhance its autonomy and experimental throughput. The authors also emphasize that the crystallinity index serves as a proxy optimization metric and does not necessarily guarantee improved porosity or water uptake.

MOFGen is a multi-agent AI system developed to accelerate the discovery of MOFs while ensuring their synthetic feasibility.[67] The framework integrates LLMs, diffusion-based generative models, machine-learning force fields, quantum mechanical computations, synthesis accessibility predictors, and robotic experimentation. Within the system, an LLM agent referred to as MOFMaster defines design constraints and coordinates the overall workflow. LinkerGen proposes novel linker molecules based on these constraints, and CrystalGen subsequently generates three-dimensional crystal structures using a diffusion model trained on MOF data. Candidate structures are then optimized using QForge, which applies quantum mechanical screening and filters out unstable or non-porous configurations. The authors also report that the generated linker chemistry exhibits features consistent with experimental trends in Zn-based MOFs, including a high prevalence of dicarboxylate and aromatic motifs and comparatively rare pyridine-containing linkers. They further observe that structures generated with Zn SBUs tended to display greater stability relative to other metal SBUs, aligning with the prevalence of Zn-based frameworks in experimental databases. The synthetic accessibility of remaining candidates is assessed by SynthABLE using a set of machine-learned rules derived from experimental data. QHarden then performs energy refinement by sequentially applying different levels of density functional theory. Finally, SynthGen conducts experimental validation through high-throughput robotic synthesis using a programmable liquid-handling platform that is guided by LLM-generated instructions and followed by X-ray characterization.

Using a combination of crossover mutation, structure reimagining, and *de novo* generation, MOFGen produced five experimentally realized MOFs, including ones incorporating a previously unused linker in MOFs, 2,3-dimethyl-2-butenedioic acid. Although human intervention was still necessary during model reliability assessments and structure revision, the system represents an early example of integrating LLM-based reasoning with autonomous experimentation. In its current form, MOFGen may be regarded as a transitional stage toward self-driving laboratories, where computational design and robotic execution are connected within a unified workflow.

As a complementary example, the study demonstrates the integration of LLMs with robotic synthesis to accelerate the discovery of greener synthetic routes for porous materials.[68] Vu *et al.* used an LLM to extract and structure nitrate-based Zn-HKUST-1 synthesis conditions and then infer plausible concentration ranges for substituting $Zn(NO_3)_2$ with $ZnCl_2$, which the authors described as an environmentally preferable alternative. Based on the data extracted by the LLM, candidate experimental conditions were identified, streamlining the traditional process of manual literature review.

These conditions were then executed using an OT-2 pipetting robot, which rapidly screened 22 concentration variations in five minutes, significantly reducing experimental setup time compared to manual handling. After synthesis, optical microscopy images were automatically evaluated using a CLIP-based classifier capable of distinguishing crystalline *versus* non-crystalline products.

Through this workflow, Zn-HKUST-1 crystals were successfully synthesized at a $ZnCl_2$ concentration of 0.6 M, and scanning electron microscopy (SEM) confirmed the expected cubic morphology. The obtained products were further verified against International Centre for Diffraction Data (ICDD) reference patterns by matching XRD peak positions and lattice constants, which were found to be close to reported literature values. While the combination of LLM reasoning, automation, and AI-based evaluation considerably reduced experiment iteration time, human oversight remained necessary, and the pipeline is not yet fully autonomous. This caution reflects the risk of relation-type hallucination, in which the AI may incorrectly predict promising synthesis conditions that are

experimentally unfeasible or suboptimal despite appearing chemically plausible. Nevertheless, this work demonstrates a meaningful progression toward self-driving laboratory frameworks by integrating multiple stages of the experimental cycle under AI guidance.

Taken together, these developments reflect a gradual shift from LLM-assisted reasoning toward increasingly automated experimentation in porous materials research. While current systems remain dependent on human oversight and are limited by model reliability and hardware constraints, they establish the groundwork for more integrated workflows. As multimodal understanding, tool interoperability, and experiment-aware feedback continue to advance, LLM-driven platforms are expected to accelerate discovery and bring the field closer to practical self-driving laboratories. At the same time, scaling these systems toward higher levels of autonomy will require careful consideration of failure modes, safety governance, and structured risk assessment frameworks, particularly when operating under potentially hazardous experimental conditions.

3. Discussion

3.1 Limitations of LLMs in porous materials research

Despite the promising progress and increasing adoption of LLMs in porous materials research, several limitations remain. One notable challenge is the continued reliance on proprietary commercial models, particularly in experimental or deployment-level settings. As shown in Table 1, GPT-4o appears most frequently across recent studies, reflecting its strong and stable performance, as also illustrated in Fig. 6. However, GPT-4o pricing remains non-trivial at $2.50 per 1 M input tokens,[126] which may become costly for workflows requiring large contextual windows, retrieval augmentation, or iterative reasoning. In practice, per-token pricing alone does not fully capture the cost of multi-step LLM workflows, as practical expenditures in porous materials research are highly task-dependent. Many recent studies employ multi-agent architectures or iterative tool-assisted reasoning, where cost scales with the number of agent calls or reasoning steps, retrieval context length, and verification cycles rather than a single inference call.[127] Recent evaluations of LLM-driven laboratory automation similarly report task-level token usage and agent interaction steps, underscoring that operational efficiency can meaningfully influence overall cost beyond raw per-token pricing. Systematic reporting of token usage and agent calls would therefore improve cost transparency and enable more meaningful comparisons across studies. Recent benchmarking efforts comparing 27 open-source and commercial models report that models such as Mistral Small 24B Instruct achieve performance approaching GPT-4o at a lower operational cost.[128] These findings suggest that open-source alternatives may help reduce cost sensitivity in future LLM-enabled research workflows. Nevertheless, adopting open-source models does not eliminate practical barriers. Deploying and maintaining local LLM systems often requires substantial computational resources, infrastructure management, and domain-specific optimization, which may exceed the capacity of many experimental laboratories. In addition, performance differences across model classes on domain-specific tasks have been reported, as illustrated in Fig. 6, suggesting that cost advantages may not always translate into equivalent analytical reliability. Finally, model outputs in both proprietary and open-source systems can exhibit sensitivity to minor prompt variations, highlighting a broader brittleness inherent to current LLM paradigms rather than a model-type-specific limitation. Beyond these practical and performance-related considerations, long-term reproducibility also warrants attention when relying on commercial LLM services.[129] Although many providers offer date-stamped model identifiers, default model aliases and backend implementations may evolve over time, underscoring the importance of precise model specification and documentation. In parallel, open-source models with fixed checkpoints provide greater transparency and environmental control, although full reproducibility in either setting still depends on consistent software and hardware configurations. Archiving prompt–response pairs used in analysis can further enhance traceability.

A second limitation is the potential for bias in LLM-generated outputs within porous materials research.[130] This issue arises in part because scientific literature is disproportionately skewed toward reporting successful experiments, while failed or inconclusive outcomes are rarely documented. Taniike and Takahashi note that most publications emphasize high-performing catalysts or successful reaction outcomes, and data-driven models may learn to regard only reported catalyst compositions or reaction conditions as correct solutions.[131] As a result, model outputs may reproduce established strategies rather than propose unconventional hypotheses or underexplored directions. More specifically, literature imbalance may affect different LLM workflows in different ways.[132] In schema-based extraction of reported synthesis conditions, performance may depend more strongly on reporting style and prompt design than on literature frequency alone. In contrast, for generative tasks such as suggesting synthesis pathways or prioritizing candidates, models may favor well-represented families and conventional routes that are more frequently observed in training data and the literature.[133] Systematic, class-stratified evaluations across porous material families remain limited and would benefit from more dedicated benchmarking efforts. In addition, unsuccessful attempts are rarely quantified in reported case studies, making it difficult to rigorously assess the true discovery efficiency of LLM-assisted workflows. Systematic reporting of failure rates and attempted conditions would provide a more realistic evaluation of model-guided synthesis and help mitigate publication bias.

Finally, achieving fully autonomous self-driving laboratories remains difficult at the current stage. For example, certain steps in porous material synthesis, such as thermal–solvation processes, still require human intervention and cannot yet be executed reliably by automated platforms.[68] In addition, while LLMs excel at natural language reasoning, they do not inherently understand causal relationships, physical constraints, or experimental feasibility. For instance, Mandal *et al.* report that

although Claude-3.5-Sonnet performs well on materials science benchmarks, its performance drops notably when deployed in an autonomous atomic force microscopy (AFM) framework.[127] They further observe that LLM-controlled experimental behavior can be highly sensitive to minor variations in prompt phrasing, introducing instability in execution. Moreover, Kitchin notes that dynamically generated experimental procedures may pose reproducibility challenges and raise safety or security concerns when executed without adequate oversight.[134] Taken together, these observations indicate that, despite rapid progress, LLMs continue to face substantial obstacles in enabling fully automated experimental platforms.

3.2 Prompt engineering *vs.* fine-tuning

Across the studies reviewed in Section 2, two dominant paradigms have emerged for adapting LLMs to porous materials research: prompt engineering and fine-tuning. In the context of LLM adaptation with prompt engineering and fine-tuning (Section 2.2), both strategies were actively explored depending on the task requirements. In contrast, studies of autonomous or self-driving laboratories (Section 2.3) relied primarily on prompt engineering because high-level reasoning, real-time decision-making, and interaction with external tools can often be handled with well-designed prompts alone, without requiring a domain-specialized model. These approaches represent distinct strategies for extending LLM capability, each offering different trade-offs in terms of cost, flexibility, scalability, and robustness in real-world deployment.

Prompt engineering relies on zero-shot[135] or few-shot[136] learning, where the model performs a task (i) without examples (zero-shot) or (ii) with a small number of in-context examples (few-shot) to shape the intended behavior without modifying model parameters. Zero-shot learning allows the model to generalize purely from prior pretraining, whereas few-shot learning leverages minimal contextual demonstrations to steer task behavior. Because no additional training is required, this approach incurs relatively low financial and computational cost and can be implemented using limited domain-specific data. Moreover, prompt-based control allows LLMs to flexibly handle a wide range of task variations, such as data extraction from literature, hypothesis generation, and iterative decision-making, by adjusting instructions rather than model weights. For example, CCA[51] and GPT-4V Image Mining[52] employed prompt-based strategies for literature data extraction and multimodal figure analysis without parameter updates, demonstrating that structured information retrieval tasks can be performed effectively through well-designed prompts alone. Similarly, autonomous and agent-based systems such as the GPT-4 Reticular Chemist,[59] and dZiner[66] relied primarily on prompt-driven reasoning frameworks to support materials design and synthesis planning, highlighting the practical viability of prompt engineering in complex, multi-step workflows. This flexibility makes prompt engineering particularly attractive for autonomous or semi-autonomous research systems, where models must respond adaptively to new experimental outcomes, shifting objectives, or unforeseen edge cases.

Fine-tuning, in contrast, involves explicitly updating model parameters using curated domain-specific datasets to create specialized LLMs optimized for a narrow class of tasks. In porous materials research, such fine-tuning substantially improves accuracy in tasks that require precise interpretation of scientific information. For example, previous work, such as L2M3,[53] Paragraph2MOFInfo,[33] and NERRE Extractor,[56] demonstrate markedly higher reliability in data extraction, categorization, and entity recognition. Moreover, fine-tuned LLMs benefit from domain-specific knowledge that can directly support design-oriented applications, such as the LLM-based hydrophobicity predictor or the MOF linker mutation model. However, this specialization comes at a cost. Fine-tuned models may lose some of the generality and creative reasoning capacity of their base counterparts, making them less suitable for open-ended or exploratory tasks. In addition, fine-tuning requires substantial investment in data collection, annotation, and quality control, as well as significant computational resources. For proprietary LLMs, the need for retraining or customized fine-tuning can be restricted or expensive, further limiting practical deployment. When domain-specific datasets are relatively limited in size, fine-tuning can enhance task-specific performance. To mitigate overfitting, Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) are particularly effective for sparse materials data.[137] Nevertheless, additional validation remains essential to ensure robustness beyond the training distribution.

Within fine-tuning approaches, parameter-efficient variants such as LoRA-based adaptation have also been explored. For example, SciToolAgent[65] fine-tuned Qwen2.5-7B using a LoRA configuration, and L2M3 (ref. 53) applied LoRA to Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct models for comparative evaluation. However, these implementations were primarily used to benchmark or contrast model performance rather than as widely deployed adaptation strategies, and reported performance remained below that of larger proprietary GPT-based systems in the respective evaluation settings. In porous materials research, parameter-efficient tuning therefore appears less frequently as a dominant operational paradigm. As reflected in Table 1, prompt engineering and full fine-tuning remain the prevailing strategies in current practice.

Taken together, prompt engineering and fine-tuning should be viewed not as competing remedies but as complementary tools within the LLM adaptation landscape. Across the reviewed studies, the choice between these strategies has been largely task-dependent. Prompt engineering favors rapid prototyping, low-cost deployment, and adaptability to dynamic research workflows, and has been predominantly adopted in settings requiring adaptive multi-step reasoning and tool interaction. In contrast, fine-tuning emphasizes precision, reproducibility, and task-specific robustness, and has been more commonly used for narrowly defined, high-precision extraction or classification tasks with standardized outputs. In porous materials research, where both exploratory creativity and reliable information

extraction are essential, the optimal strategy often lies in combining these approaches.

3.3 Effect of temperature on reliability of LLM-driven workflows

The temperature parameter in LLMs controls the degree of diversity and creativity in generated outputs. Statistically, it originates from the Boltzmann distribution, where temperature rescales the logits before sampling.

$$P_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

A lower temperature sharpens the probability distribution, making the model increasingly deterministic, whereas a higher temperature flattens the distribution and increases diversity. Most LLMs default to a temperature of around 1.0. In the OpenAI API documentation, lower values (*e.g.*, 0.2) are described as producing more focused and deterministic outputs, whereas higher values (*e.g.*, 0.8) lead to greater randomness.[138] In the porous materials studies reviewed here, temperatures in the range of 0.0–0.3 are typically used when accuracy and stability are critical, while values above 1.0 are employed for creative ideation tasks. Thus, selecting an appropriate temperature is essential for aligning model behavior with the objective of a given task.

In porous materials research, temperature settings are often task-dependent. Deterministic outputs are desirable for data extraction, entity recognition, and classification, whereas proposing novel synthesis routes or generating molecular candidates may require more diverse and exploratory outputs. This distinction is reflected in the case studies reviewed in Section 2. For instance, L2M3 and NERRE (Section 2.2) both employed a temperature of 0, ensuring consistent extraction of chemical entities and synthesis conditions from the literature.

By contrast, studies involving autonomous or generative workflows (Section 2.3) adopted a broader range of temperatures. In zeolite synthesis tasks, temperature values between 0.7 and 1.1 were used to generate diverse OSDA molecules.[64] Meanwhile, in systems like dZiner and MOFSyn, temperature values were kept low (0.3 and 0.1, respectively) to minimize hallucinations and maintain reliability during generation.

Taken together, current applications suggest that LLMs in porous materials research are generally operated at relatively deterministic settings, especially when chemical accuracy, reproducibility, and safe autonomous execution are required. This trend reflects the current priority of reliability over creativity in most real-world workflows, even though higher-temperature sampling remains valuable for creative generative tasks.

3.4 Safety considerations for LLM-driven agents and laboratory automation

As LLM-driven agents are increasingly incorporated into porous materials research workflows, safety must be addressed alongside performance, optimization speed, and experimental throughput. Although laboratory automation can reduce routine human exposure to hazardous reagents, recent analyses of self-driving laboratories note that increasing autonomy introduces tighter coupling between model reasoning, robotic actuation, and networked data infrastructures, thereby amplifying the consequences of decision errors.[104] As systems transition from scripted automation to adaptive autonomy, model-generated instructions may directly determine reagent selection, reaction conditions, motion trajectories, and execution timing, creating a growing safety gap between system capability and safety governance. In this setting, risks extend beyond conventional chemical hazards to include unsafe experimental decisions at the planning stage, robotic execution failures such as spills, breakage, or manipulation errors, infrastructure-level vulnerabilities including cyber-security threats and malicious misuse through dual use or jailbreak exploits, and cognitive failure modes intrinsic to LLMs such as prompt sensitivity, cumulative error propagation across multi-step workflows, misalignment of safety priorities, and hallucination.[139,140]

To mitigate these risks and operationalize epistemic humility, several concrete strategies can be employed. These include self-consistency or multi-sampling to estimate output variance, retrieval-grounded generation with traceable evidence, and tool-augmented cross-checking such as ReAct.[141] Rather than issuing unconditional recommendations, such mechanisms allow models to communicate confidence levels and flag high-uncertainty decisions for human oversight, which is essential for trustworthy autonomous experimentation.

The importance of structured safety validation is underscored by benchmarking efforts such as LabSafety Bench, which evaluates LLMs across scenario-based tasks including hazard identification, consequence assessment, and safe protocol recommendation.[139] In these evaluations, many advanced models achieved average scores in the 60–70% range, with notable variability across task categories and frequent failures in complex multi-factor safety scenarios. The study reports recurrent cognitive errors, including misalignment of safety priorities in which models emphasized obvious risks such as fire while overlooking more severe toxic gas release scenarios, hallucinated chemical interactions lacking mechanistic basis, and incomplete recognition of compound hazard interactions within experimental protocols. These findings demonstrate that coherent natural-language reasoning does not guarantee accurate hazard ranking or consequence prediction, underscoring the need for domain-specific safety benchmarking and validation prior to experimental deployment.

In porous-material case studies reviewed in Section 2.3, safety considerations are present but not yet formalized as analytical pillars. In the ChatGPT Research Group example, safety guidance was implemented through a chemistry consultant agent that advised on microwave irradiation, pressure buildup, and chemical hazards.[24] However, structured failure-mode classification or quantitative safety auditing was not central to the study design. MOFGen emphasized synthetic accessibility and structural validation, while the Zn-HKUST-1 robotic workflow focused on experimental acceleration and qualitative environmental reasoning.[67,68] These examples

indicate that safety components are present but not yet formalized as standalone safety architectures within porous-material SDL implementations.

To enhance the safety of LLM-integrated laboratory systems, recent studies argue that safeguards must intervene across reasoning, execution, and governance rather than relying solely on physical containment. The notion of cognitive safety proposes that experimental plans generated by a primary LLM be automatically screened prior to execution for hazards such as excessive pressure buildup, incompatible reagents, runaway exothermic conditions, or violation of predefined safety constraints, thereby reducing execution-stage hazards arising from flawed reasoning. Execution-level protection incorporates sensor-aware robotics, including vision systems that detect transparent glassware or human intrusion into shared workspaces and thermal imaging to flag abnormal heat signatures, alongside constrained motion planning that limits transfer velocities, pouring angles, and collision trajectories to reduce spills and mechanical impact. Because serious laboratory accidents are rare, digital twin environments are proposed to simulate unsafe mixing conditions, collision events, or equipment failures prior to deployment, enabling stress testing under limited empirical accident data. Governance mechanisms complement these safeguards through structured risk assessment based on likelihood, severity, system complexity, and autonomy level, traceable logging of AI-generated decisions for accountability, near-miss reporting practices, and alignment with frameworks such as the EU AI Act, ISO 42001, and transparency standards including PRISMA-AI. For porous materials research, where solvothermal synthesis in sealed autoclaves and high-pressure or microwave-assisted reactions are common, integrating such layered safety architectures will be increasingly important as LLM-driven experimentation advances toward greater autonomy. Future work in LLM-driven porous materials research should integrate safety evaluation and autonomy-level risk auditing alongside performance metrics.

4. Conclusions and perspective

In this review, we examined how LLMs are currently being applied to porous materials research, drawing on representative studies across text mining, LLM adaptation, and autonomous experimentation. We first outlined the foundations of LLMs, including their operational principles, commonly used models, and approaches for adapting them to scientific tasks through prompt engineering and fine-tuning. We then reviewed three major application domains. The first involved how natural language processing techniques were used before the advent of modern LLMs for text preprocessing, text representation, and information extraction in porous materials research. The second explored how contemporary studies leverage LLMs through prompt engineering or fine-tuning to perform data mining, structured information extraction, and domain-specific scientific applications. The third focused on the construction of human-in-the-loop and fully autonomous systems, illustrating how LLMs increasingly mediate experimental decision-making and interaction with laboratory tools.

Despite these significant advances, our discussion highlighted several remaining limitations, including high model costs, data imbalance in the scientific literature, the limited robustness of current autonomous laboratory frameworks, and the need for structured safety validation. We also discussed the trade-offs between prompt engineering and fine-tuning strategies, and the influence of temperature settings on the determinism, stability, and creativity of model outputs.

Looking ahead, several promising directions are emerging. Beyond model-level innovation, advancing LLM-driven porous materials research will require shared community infrastructure, including standardized benchmarks, curated evaluation datasets, reproducible safety guidelines, and evaluation frameworks capable of assessing genuine scientific novelty rather than performance on narrow task-specific metrics. Multimodal LLMs capable of jointly reasoning over experimental text, molecular structures, and image data like X-ray diffraction (XRD) are expected to expand the scope of tasks that may be more robustly supported through automation. In parallel, integrating LLMs with structured knowledge sources such as knowledge graphs (KG) may provide pathways toward more interpretable and constraint-aware reasoning. By retrieving relevant subgraphs rather than noisy text chunks, this KG enables the model to perform complex, multi-hop queries, such as filtering materials by precursors, application, and stability simultaneously, while promoting responses that are more transparently grounded in traceable, literature-derived evidence. This integration can help mitigate naming ambiguities by resolving coreferences to correct crystal structures, providing factually accurate answers as verified by expert evaluations. LLMs are also expected to progress from assisting as experiment planners to operating as higher-level supervisors within self-driving laboratory ecosystems. Ensuring chemical validity, minimizing hallucinations, and maintaining operational and ethical safety will be critical for this transition, highlighting the need for structured failure-mode analysis, domain-specific safety benchmarking, supervisory monitoring architectures, and accountable governance mechanisms in LLM-driven workflows. Real-world deployment of such systems will further require interoperable laboratory infrastructure, including hardware standardization, integration with laboratory information management systems (LIMS), implementation of physical safety interlocks, and alignment with emerging laboratory automation standards such as Synthetic Procedure Language (SPL), alongside appropriate regulatory and liability frameworks.

With continued advances across these areas, LLMs have the potential to evolve from supportive computational assistants into enabling technologies that contribute to increasingly autonomous and scientifically reliable self-driving laboratory ecosystems for porous materials research.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Conflicts of interest

The authors declare no competing financial interest.

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 project was supported by National Research Foundation of Korea (NRF) under grant No. RS-2024-00337004 and RS-2024-00451160.

References

1 H. Li, M. Eddaoudi, M. O'Keeffe and O. M. Yaghi, Design and synthesis of an exceptionally stable and highly porous metal-organic framework, *Nature*, 1999, **402**, 276–279.
2 U. Díaz and A. Corma, Ordered covalent organic frameworks, COFs and PAFs. From preparation to application, *Coord. Chem. Rev.*, 2016, **311**, 85–124.
3 T. K. Maji and S. Kitagawa, Chemistry of porous coordination polymers, *Pure Appl. Chem.*, 2007, **79**, 2155–2177.
4 N. Kordala and M. Wyszkowski, Zeolite properties, methods of synthesis, and selected applications, *Molecules*, 2024, **29**, 1069.
5 M. Ni, L. Zhou, Y. Liu and R. Ni, Advances in the synthesis and applications of porous carbon materials, *Front. Chem.*, 2023, **11**, 1205280.
6 D. Ma, Y. Wang, A. Liu, S. Li, C. Lu and C. Chen, Covalent organic frameworks: Promising materials as heterogeneous catalysts for CC bond formations, *Catalysts*, 2018, **8**, 404.
7 Y. Chen, W. Lu, M. Schroder and S. Yang, Analysis and refinement of host–guest interactions in metal–organic frameworks, *Acc. Chem. Res.*, 2023, **56**, 2569–2581.
8 R. Singh, L. Wang, K. Ostrikov and J. Huang, Designing carbon-based porous materials for carbon dioxide capture, *Adv. Mater. Interfaces*, 2024, **11**, 2202290.
9 B. Petrovic, M. Gorbounov and S. M. Soltani, Influence of surface modification on selective CO2 adsorption: A technical review on mechanisms and methods, *Microporous Mesoporous Mater.*, 2021, **312**, 110751.
10 A. A. Lahcen, S. G. Surya, T. Beduk, M. T. Vijjapu, A. Lamaoui, C. Durmus, S. Timur, O. Shekhah, V. Mani and A. Amine, Metal–organic frameworks meet molecularly imprinted polymers: insights and prospects for sensor applications, *ACS Appl. Mater. Interfaces*, 2022, **14**, 49399–49424.
11 A. Maleki, M. A. Shahbazi, V. Alinezhad and H. A. Santos, The progress and prospect of zeolitic imidazolate frameworks in cancer therapy, antibacterial activity, and biomineralization, *Adv. Healthcare Mater.*, 2020, **9**, 2000248.
12 Y. Kim, W. Lee and J. Kim, Amorphous Metal–Organic Framework Database for Amorphization Prediction and CO2 Direct Air Capture Screening, *ACS Appl. Mater. Interfaces*, 2025, **17**, 49647–49659.
13 H. Kum and J. Kim, High-Throughput Screening of Ru-Based MOF-Supported Single-Atom Catalysts for Hydrogen Evolution Reaction *Via* Machine Learning Interatomic Potential, *ACS Catal.*, 2025, **15**, 19756–19767.
14 S. Han and J. Kim, Design and screening of metal–organic frameworks for ethane/ethylene separation, *ACS Omega*, 2023, **8**, 4278–4284.
15 I. G. Clayson, D. Hewitt, M. Hutereau, T. Pope and B. Slater, High throughput methods in the synthesis, characterization, and optimization of porous materials, *Adv. Mater.*, 2020, **32**, 2002780.
16 A. M. Mroz, V. Posligua, A. Tarzia, E. H. Wolpert and K. E. Jelfs, Into the unknown: how computation can help explore uncharted material space, *J. Am. Chem. Soc.*, 2022, **144**, 18730–18743.
17 C. Duan, A. Nandy and H. J. Kulik, Machine learning for the discovery, design, and engineering of materials, *Annu. Rev. Chem. Biomol. Eng.*, 2022, **13**, 405–429.
18 J. Park, H. Kim, Y. Kang, Y. Lim and J. Kim, From data to discovery: recent trends of machine learning in metal–organic frameworks, *JACS Au*, 2024, **4**, 3727–3743.
19 W. Wang, X. Jiang, S. Tian, P. Liu, D. Dang, Y. Su, T. Lookman and J. Xie, Automated pipeline for superalloy data by text mining, *npj Comput. Mater.*, 2022, **8**, 9.
20 H. Park, Y. Kang, W. Choe and J. Kim, Mining insights on metal–organic framework synthesis from scientific literature texts, *J. Chem. Inf. Model.*, 2022, **62**, 1190–1198.
21 J. Zhang, X. Chen, X. Ye, Y. Yang and B. Ai, Large Language Model in Materials Science: Roles, Challenges, and Strategic Outlook, *Adv. Intell. Syst.*, 2025, 202500085.
22 S. Bae, M. Jeon and H. R. Moon, Text Mining in MOF Research: From Manual Curation to Large Language Model-Based Automation, *Chem. Commun.*, 2025, **61**(60), 11083–11094.
23 Y. Kang and J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, *Nat. Commun.*, 2024, **15**, 4705.
24 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, ChatGPT research group for optimizing the crystallinity of MOFs and COFs, *ACS Cent. Sci.*, 2023, **9**, 2161–2170.
25 M. C. Ramos, C. J. Collison and A. D. White, A review of large language models and autonomous agents in chemistry, *Chem. Sci.*, 2025, **16**(6), 2514–2572.
26 L. Reynolds and K. McDonell, Prompt programming for large language models: Beyond the few-shot paradigm, in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–7.
27 J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le and D. Zhou, Chain-of-thought prompting elicits

reasoning in large language models, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 24824–24837.
28 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih and T. Rocktäschel, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 9459–9474.
29 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems*, 2017, vol. 30.
30 Y. Kang, H. Park, B. Smit and J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks, *Nat. Mach. Intell.*, 2023, **5**, 309–318.
31 S. Han, Y. Kang, H. Park, J. Yi, G. Park and J. Kim, Multimodal transformer for property prediction in polymers, *ACS Appl. Mater. Interfaces*, 2024, **16**, 16853–16860.
32 S. Kamnis and K. Delibasis, High entropy alloy property predictions using a transformer-based language model, *Sci. Rep.*, 2025, **15**, 11861.
33 W. Zhang, Q. G. Wang, X. T. Kong, J. C. Xiong, S. K. Ni, D. H. Cao, B. Y. Niu, M. G. Chen, Y. M. Li, R. Z. Zhang, Y. T. Wang, L. H. Zhang, X. T. Li, Z. P. Xiong, Q. Shi, Z. M. Huang, Z. Y. Fu and M. Y. Zheng, Fine-tuning large language models for chemical text mining, *Chem. Sci.*, 2024, **15**, 10600–10611.
34 Y. Song, S. Miret, H. Zhang and B. Liu, HoneyBee: Progressive instruction finetuning of large language models for materials science, in *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 5724–5739.
35 Z. Cao and L. Wang, CrystalFormer-RL: Reinforcement Fine-Tuning for Materials Design, *arXiv*, 2025, preprint, arXiv:2504.02367, DOI: 10.48550/arXiv.2504.02367.
36 S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz and J. Weston, in *Findings of the association for computational linguistics: ACL*, 2024, pp. 3563–3578.
37 F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang and F. Meng, Towards large reasoning models: A survey of reinforced reasoning with large language models, *Patterns*, 2025, **6**, 101370.
38 J. Choi and B. Lee, Accelerated materials language processing enabled by GPT, *arXiv*, 2023, preprint, arXiv:2308.09354, DOI: 10.48550/arXiv.2308.09354.
39 G. Lei, R. Docherty and S. J. Cooper, Materials science in the era of large language models: a perspective, *Digital Discovery*, 2024, **3**, 1257–1272.
40 M. P. Polak and D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, *Nat. Commun.*, 2024, **15**, 1569.
41 J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman and S. Anadkat, Gpt-4 technical report, *arXiv*, 2023, preprint, arXiv:2303.08774, DOI: 10.48550/arXiv.2303.08774.
42 H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro and F. Azhar, Llama: Open and efficient foundation language models, *arXiv*, 2023, preprint, arXiv:2302.13971, DOI: 10.48550/arXiv.2302.13971.
43 B. A. Nvidia, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, A. Brundyn, J. Casper, B. Catanzaro, S. Clay and J. Cohen, Nemotron-4 340b technical report, *arXiv*, 2024, preprint, arXiv:2406.11704, DOI: 10.48550/arXiv.2406.11704.
44 D. S. Chaplot, and A. q. Jiang, Alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed, *arXiv*, 2023, preprint, arXiv:2310.06825, DOI: 10.48550/arXiv.2310.06825, 3.
45 T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng and H. Zhao, Chatglm: A family of large language models from glm-130b to glm-4 all tools, *arXiv*, 2024, preprint, arXiv:2406.12793, DOI: 10.48550/arXiv.2406.12793.
46 E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay and Q. Malartic, The falcon series of open language models, *arXiv*, 2023, preprint, arXiv:2311.16867, DOI: 10.48550/arXiv.2311.16867.
47 X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du and Z. Fu, Deepseek llm: Scaling open-source language models with longtermism, *arXiv*, 2024, preprint, arXiv:2401.02954, DOI: 10.48550/arXiv.2401.02954.
48 Anthropic, *Claude 4.5 Sonnet and Opus - Anthropic Large Language Models*, 2025, https://platform.claude.com/docs/en/resources/overview.
49 A. Liu, B. Feng, B. Xue, B. Wang, B. Wu and C. Lu, *et al.*, Deepseek-v3 technical report, *arXiv*, 2024, preprint arXiv:2412.19437, DOI: 10.48550/arXiv.2412.19437.
50 J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou and J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. CoRR, abs/2308.12966, *arXiv*, 2023, preprint, arXiv.2308.12966, DOI: 10.48550/arXiv.2308.12966.
51 Z. L. Zheng, O. F. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
52 Z. L. Zheng, Z. G. He, O. Khattab, N. Rampal, M. A. Zaharia, C. Borgs, J. T. Chayes and O. M. Yaghi, Image and data mining in reticular chemistry powered by GPT-4V, *Digital Discovery*, 2024, **3**, 491–501.
53 Y. Kang, W. Lee, T. Bae, S. Han, H. Jang and J. Kim, Harnessing Large Language Models to Collect and Analyze Metal-Organic Framework Property Data Set, *J. Am. Chem. Soc.*, 2025, **147**, 3943–3958.

54 L. Hu, Z. R. Zhou and G. Z. Jia, A one-shot automated framework based on large language model and AutoML: Accelerating the design of porous carbon materials and carbon capture optimization, *Sep. Purif. Technol.*, 2025, 376.
55 J. Zhao, M. Yan, Z. Shi, C. Zhao, L. Qi, Q. Hao and H. Liu, Prediction of COF Synthesis Reaction Conditions Based on Data Mining and Machine Learning, *ACS Appl. Polym. Mater.*, 2025, **7**, 16973–16981.
56 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, Structured information extraction from scientific text with large language models, *Nat. Commun.*, 2024, **15**(1), 1418.
57 M. Ansari and S. M. Moosavi, Agent-based learning of materials datasets from the scientific literature, *Digital Discovery*, 2024, **3**, 2607–2617.
58 X. F. Bai, Y. B. Xie, X. Zhang, H. G. Han and J. R. Li, Evaluation of Open-Source Large Language Models for Metal-Organic Frameworks Research, *J. Chem. Inf. Model.*, 2024, **64**, 4958–4965.
59 N. Rampal, K. Y. Wang, M. Burigana, L. X. Hou, J. Al-Johani, A. Sackmann, H. S. Murayshid, W. A. AlSumari, A. M. AlAbdulkarim, N. E. Alhazmi, M. O. Alawad, C. Borgs, J. T. Chayes and O. M. Yaghi, Single and Multi-Hop Question-Answering Datasets for Reticular Chemistry with GPT-4-Turbo, *J. Chem. Theory Comput.*, 2024, **20**, 9128–9137.
60 X. Y. Wu and J. W. Jiang, Can large language models predict the hydrophobicity of metal-organic frameworks?, *J. Mater. Chem. A*, 2025, **13**(25), 19307–19315.
61 Z. L. Zheng, A. H. Alawadhi, S. Chheda, S. E. Neumann, N. Rampal, S. C. Liu, H. Nguyen, Y. H. Lin, Z. C. Rong, J. I. Siepmann, L. Gagliardi, A. Anandkumar, C. Borgs, J. T. Chayes and O. M. Yaghi, Shaping the Water-Harvesting Behavior of Metal-Organic Frameworks Aided by Fine-Tuned GPT Models, *J. Am. Chem. Soc.*, 2023, **145**, 28284–28295.
62 Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes and O. M. Yaghi, A GPT-4 reticular chemist for guiding MOF discovery, *Angew. Chem., Int. Ed.*, 2023, **62**, e202311983.
63 J. Lin, D. Zhao, S. Lu, R. Li, X. Xu, Z. Wang, W. Li, Y. Ji, C. Zhang and L. Shi, Conversational Large-Language-Model Artificial Intelligence Agent for Accelerated Synthesis of Metal–Organic Frameworks Catalysts in Olefin Hydrogenation, *ACS Nano*, 2025, **19**(26), 23840–23858.
64 S. Ito, K. Muraoka and A. Nakayama, Knowledge-Informed Molecular Design for Zeolite Synthesis Using General-Purpose Pretrained Large Language Models Toward Human-Machine Collaboration, *Chem. Mater.*, 2025, **37**, 2447–2456.
65 K. Ding, J. Yu, J. Huang, Y. Yang, Q. Zhang and H. Chen, SciToolAgent: a knowledge-graph-driven scientific agent for multitool integration, *Nat. Comput. Sci.*, 2025, 1–11.
66 M. Ansari, J. Watchorn, C. E. Brown and J. S. Brown, dZiner: Rational inverse design of materials with ai agents, *arXiv*, 2024, preprint, arXiv:2410.03963, DOI: 10.48550/arXiv.2410.03963.
67 T. J. Inizan, S. Yang, A. Kaplan, Y.-h. Lin, J. Yin, S. Mirzaei, M. Abdelgaid, A. H. Alawadhi, K. Cho and Z. Zheng, System of agentic AI for the discovery of metal-organic frameworks, *arXiv*, 2025, preprint, arXiv:2504.14110, DOI: 10.48550/arXiv.2504.14110.
68 V.-H. Vu, K.-H. Bui, K. D. Dang, M. Duong-Tuan, D. D. Le and T. Nguyen-Dang, Finding environmental-friendly chemical synthesis with AI and high-throughput robotics, *J. Sci. Adv. Mater. Devices*, 2025, **10**, 100818.
69 T. Gupta, M. Zaki and N. M. A. Krishnan, Mausam, MatSciBERT: A materials domain language model for text mining and information extraction, *npj Comput. Mater.*, 2022, **8**, 102.
70 S. Huang and J. M. Cole, BatteryDataExtractor: battery-aware text-mining software embedded with BERT models, *Chem. Sci.*, 2022, **13**, 11487–11495.
71 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, ChemicalTagger: A tool for semantic text-mining in chemistry, *J. Cheminf.*, 2011, **3**, 17.
72 P. Qi, Y. H. Zhang, Y. H. Zhang, J. Bolton and C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, *58th Annual Meeting of the Association for Computational Linguistics (Acl 2020): System Demonstrations*, 2020, pp. 101–108.
73 S. Zhao and N. Birbilis, Searching for chromate replacements using natural language processing and machine learning algorithms, *npj Mater. Degrad.*, 2023, **7**, 2.
74 J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019)*, vol. 1, 2019, pp. 4171–4186.
75 Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. D. Liu, T. Naumann, J. F. Gao and H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, *ACM Trans. Comput. Heal.*, 2022, **3**(1), 1–23.
76 Z. Huang, W. Xu and K. Yu, Bidirectional LSTM-CRF models for sequence tagging, *arXiv*, 2015, preprint, arXiv:1508.01991, DOI: 10.48550/arXiv.1508.01991.
77 M. C. Swain and J. M. Cole, ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
78 P. Shetty, A. C. Rajan, C. Kuenneth, S. Gupta, L. P. Panchumarti, L. Holm, C. Zhang and R. Ramprasad, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, *npj Comput. Mater.*, 2023, **9**, 52.
79 O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan and G. Ceder, Text-mined dataset of inorganic materials synthesis recipes, *Sci. Data*, 2019, **6**, 203.

80 O. Tayfuroglu, A. Kocak and Y. Zorlu, In Silico Investigation into H2 Uptake in MOFs: Combined Text/Data Mining and Structural Calculations, *Langmuir*, 2020, **36**, 119–129.
81 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, The Cambridge Structural Database, *Acta Crystallogr., Sect. B:Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.
82 T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza and M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials, *Microporous Mesoporous Mater.*, 2012, **149**, 134–141.
83 L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S. M. Moosavi, J. L. Cordiner, J. C. Cole and P. Z. Moghadam, DigiMOF: A Database of Metal–Organic Framework Synthesis Information Generated *via* Text Mining, *Chem. Mater.*, 2023, **35**, 4510–4524.
84 L. Zoubritzky and F.-X. Coudert, CrystalNets.jl: Identification of Crystal Topologies, *SciPost Chem.*, 2022, **1**, 005.
85 E. Pan, S. Kwon, Z. Jensen, M. Xie, R. Gómez-Bombarelli, M. Moliner, Y. Román-Leshkov and E. Olivetti, ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters, *ACS Cent. Sci.*, 2024, **10**, 729–743.
86 S. He, W. Du, X. Peng and X. Li, ZeoReader: Automated extraction of synthesis steps from zeolite synthesis literature for autonomous experiments, *Chem. Eng. Sci.*, 2025, **302**, 120916.
87 Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich and M. Tsotsalas, MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning, *Angew. Chem., Int. Ed.*, 2022, **61**, e202200242.
88 Y. G. Chung, E. Haldoupis, B. J. Bucior, M. Haranczyk, S. Lee, H. Zhang, K. D. Vogiatzis, M. Milisavljevic, S. Ling, J. S. Camp, B. Slater, J. I. Siepmann, D. S. Sholl and R. Q. Snurr, Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019, *J. Chem. Eng. Data*, 2019, **64**, 5985–5998.
89 A. Nandy, G. Terrones, N. Arunachalam, C. Duan, D. W. Kastner and H. J. Kulik, MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks, *Sci. Data*, 2022, **9**, 74.
90 J. P. Janet and H. J. Kulik, Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships, *J. Phys. Chem.*, 2017, **121**, 8939–8954.
91 Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner and E. A. Olivetti, Discovering Relationships between OSDAs and Zeolites through Data Mining and Generative Neural Networks, *ACS Cent. Sci.*, 2021, **7**, 858–867.
92 R. Todeschini, P. Gramatica, R. Provenzani and E. Marengo, Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons, *Chemom. Intell. Lab. Syst.*, 1995, **27**, 221–229.
93 R. Todeschini and P. Gramatica, The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling, *SAR QSAR Environ. Res.*, 1997, **7**, 89–115.
94 K. Hira, M. Zaki, D. Sheth, Mausam and N. M. A. Krishnan, Reconstructing the materials tetrahedron: challenges in materials information extraction, *Digital Discovery*, 2024, **3**, 1021–1037.
95 J. Choi and B. Lee, Accelerating materials language processing with large language models, *Commun. Mater.*, 2024, **5**(1), 13.
96 Y. Z. Zheng, H. Y. Koh, J. X. Ju, A. T. N. Nguyen, L. T. May, G. I. Webb and S. R. Pan, Large language models for scientific discovery in molecular property prediction, *Nat. Mach. Intell.*, 2025, **7**(3), 437–447.
97 S. Kim, Y. Jung and J. Schrier, Large Language Models for Inorganic Synthesis Predictions, *J. Am. Chem. Soc.*, 2024, **146**, 19654–19659.
98 S. Kim, J. Schrier and Y. Jung, Explainable Synthesizability Prediction of Inorganic Crystal Polymorphs Using Large Language Models, *Angew. Chem., Int. Ed.*, 2025, 64.
99 N. Alampara, M. Schilling-Wilhelmi, M. Ríos-García, I. Mandal, P. Khetarpal, H. S. Grover, N. M. A. Krishnan and K. M. Jablonka, Probing the limitations of multimodal language models for chemistry and materials research, *Nat. Comput. Sci.*, 2025, **5**, 952–961.
100 O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi and H. Moazam, Dspy: Compiling declarative language model calls into self-improving pipelines, *arXiv*, 2023, preprint, arXiv:2310.03714, DOI: 10.48550/arXiv.2310.03714.
101 Y. Su, X. Wang, Y. Ye, Y. Xie, Y. Xu, Y. Jiang and C. Wang, Automation and machine learning augmented by large language models in a catalysis study, *Chem. Sci.*, 2024, **15**, 12200–12233.
102 N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim and G. Ceder, Toward autonomous design and synthesis of novel inorganic materials, *Mater. Horiz.*, 2021, **8**, 2169–2198.
103 W. Yuan, G. Chen, Z. Wang and F. You, Empowering Generalist Material Intelligence with Large Language Models, *Adv. Mater.*, 2025, 2502771.
104 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson and M. Skreta, Self-driving laboratories for chemistry and materials science, *Chem. Rev.*, 2024, **124**, 9633–9732.
105 R. S. Gadde, S. Devaguptam, F. Ren, R. Mittal, L. Dong, Y. Wang and F. Liu, Chatbot-assisted quantum chemistry for explicitly solvated molecules, *Chem. Sci.*, 2025, **16**, 3852–3864.
106 G. Inc., Conversational Agents (Dialogflow CX) Documentation, https://docs.cloud.google.com/dialogflow/cx/docs?hl=ko.
107 Y. Zou, A. H. Cheng, A. Aldossary, J. Bai, S. X. Leong, J. A. Campos-Gonzalez-Angulo, C. Choi, C. T. Ser, G. Tom and A. Wang, El Agente: An autonomous agent for quantum chemistry, *Matter*, 2025, **8**(7), 102263.

108 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang and G. Zou, A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand, *J. Am. Chem. Soc.*, 2025, **147**, 12534–12545.

109 L. Huang, C. Zhang, Y. Fu, Y. Jiang, E. He, M.-Q. Qi, M.-H. Du, X.-J. Kong, J. Cheng and L. Cronin, Natural-Language-Interfaced Robotic Synthesis for AI-Copilot-Assisted Exploration of Inorganic Materials, *J. Am. Chem. Soc.*, 2025, **147**(26), 23014–23025.

110 J. Johnson, M. Douze and H. Jégou, Billion-scale similarity search with GPUs, *IEEE Trans. Big Data.*, 2019, **7**, 535–547.

111 S. Robertson and H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, *Found. Trends Inf. Retr.*, 2009, **3**, 333–389.

112 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.*, 2024, **6**, 161–169.

113 S. Es, J. James, L. E. Anke and S. Schockaert, in *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, 2024, pp. 150–158.

114 F. Daeyaert and M. W. Deem, Design of organic structure directing agents to control the synthesis of zeolites for carbon capture and storage, *RSC Adv.*, 2019, **9**, 41934–41942.

115 J. K. Lee, J. Shin, N. H. Ahn, A. Turrina, M. B. Park, Y. Byun, S. J. Cho, P. A. Wright and S. B. Hong, A Family of Molecular Sieves Containing Framework-Bound Organic Structure-Directing Agents, *Angew. Chem.*, 2015, **127**, 11249–11253.

116 D. Schwalbe-Koda and R. Gómez-Bombarelli, Supramolecular recognition in crystalline nanocavities through Monte Carlo and Voronoi network algorithms, *J. Phys. Chem. C*, 2021, **125**, 3009–3017.

117 J. E. Schmidt, M. W. Deem, C. Lew and T. M. Davis, Computationally-guided synthesis of the 8-ring zeolite AEI, *Top. Catal.*, 2015, **58**, 410–415.

118 P. Wagner, Y. Nakagawa, G. S. Lee, M. E. Davis, S. Elomari, R. C. Medrud and S. Zones, Guest/host relationships in the synthesis of the novel cage-based zeolites SSZ-35, SSZ-36, and SSZ-39, *J. Am. Chem. Soc.*, 2000, **122**, 263–273.

119 D. Dubbeldam, S. Calero, D. E. Ellis and R. Q. Snurr, RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials, *Mol. Simul.*, 2016, **42**, 81–101.

120 Z. Cao, R. Magar, Y. Wang and A. Barati Farimani, Moformer: self-supervised transformer model for metal–organic framework property prediction, *J. Am. Chem. Soc.*, 2023, **145**, 2958–2967.

121 A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein and R. Q. Snurr, Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery, *Matter*, 2021, **4**, 1578–1597.

122 S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan and Y. Cao, in *The eleventh international conference on learning representations*, 2022.

123 E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine and K. Leyton-Brown, MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, External knowledge sources and discrete reasoning, *arXiv*, 2022, preprint, arXiv:2205.00445, DOI: 10.48550/arXiv.2205.00445.

124 Y. Xie, C. Zhang, H. Deng, B. Zheng, J.-W. Su, K. Shutt and J. Lin, Accelerate synthesis of metal–organic frameworks by a robotic platform and bayesian optimization, *ACS Appl. Mater. Interfaces*, 2021, **13**, 53485–53491.

125 L. Pilz, C. Natzeck, J. Wohlgemuth, N. Scheuermann, P. G. Weidler, I. Wagner, C. Wöll and M. Tsotsalas, Fully automated optimization of robot-based MOF thin film growth *via* machine learning approaches, *Adv. Mater. Interfaces*, 2023, **10**, 2201771.

126 GPT-4 pricing, https://platform.openai.com/docs/pricing?utm_source=chatgpt.com.

127 I. Mandal, J. Soni, M. Zaki, M. M. Smedskjaer, K. Wondraczek, L. Wondraczek, N. N. Gosvami and N. A. Krishnan, Evaluating large language model agents for automation of atomic force microscopy, *Nat. Commun.*, 2025, **16**, 9104.

128 C. Caminha, M. d. L. M. Silva, I. C. Chaves, F. T. Brito, V. A. Farias and J. C. Machado, Evaluating LLMs and Prompting Strategies for Automated Hardware Diagnosis from Textual User-Reports, *arXiv*, 2025, preprint, arXiv:2507.00742, DOI: 10.48550/arXiv.2507.00742.

129 L. Chen, M. Zaharia and J. Zou, How is ChatGPT's behavior changing over time?, *Harv. Data Sci. Rev.*, 2024, **6**(2), DOI: 10.1162/99608f92.5317da47.

130 E. Xie, X. Wang, J. I. Siepmann, H. Chen and R. Q. Snurr, Generative AI for design of nanoporous materials: review and future prospects, *Digital Discovery*, 2025, **4**(9), 2336–2363.

131 T. Taniike and K. Takahashi, The value of negative results in data-driven catalysis research, *Nat. Catal.*, 2023, **6**, 108–111.

132 I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang and N. K. Ahmed, Bias and fairness in large language models: A survey, *Comput. Linguist.*, 2024, **50**, 1097–1179.

133 A. Algaba, C. Mazijn, V. Holst, F. Tori, S. Wenmackers and V. Ginis, in *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 6829–6864.

134 J. R. Kitchin, The evolving role of programming and LLMs in the development of self-driving laboratories, *APL Mach. Learn.*, 2025, **3**, 026111.

135 T. Kojima, S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa, Large language models are zero-shot reasoners, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 22199–22213.

136 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry and A. Askell, Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 1877–1901.

137 M. Szép, D. Rueckert, R. von Eisenhart-Rothe and F. Hinterwimmer, A practical guide to fine-tuning

language models with limited data, *arXiv*, preprint, arXiv:2411.09539, 2024, DOI: 10.48550/arXiv.2411.09539.

138 OpenAI API, https://developers.openai.com/api/reference/resources/completions/methods/create/?utm_source=chatgpt.com.

139 Y. Zhou, J. Yang, Y. Huang, K. Guo, Z. Emory, B. Ghosh, A. Bedar, S. Shekar, Z. Liang and P.-Y. Chen, Benchmarking large language models on safety risks in scientific laboratories, *Nat. Mach. Intell.*, 2026, 1–12.

140 S. X. Leong, C. E. Griesbach, R. Zhang, K. Darvish, Y. Zhao, A. Mandal, Y. Zou, H. Hao, V. Bernales and A. Aspuru-Guzik, Steering towards safe self-driving laboratories, *Nat. Rev. Chem.*, 2025, **9**, 707–722.

141 X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, *arXiv*, 2022, preprint, arXiv:2203.11171, DOI: 10.48550/arXiv.2203.11171.