

LLM-based MOFs Synthesis Condition Extraction using Few-Shot Demonstrations

Lei Shi^{1†}, Zhimeng Liu^{2†}, Yi Yang¹, Weize Wu¹, Yuyang Zhang¹, Hongbo Zhang³
Jing Lin², Siyu Wu¹, Zihan Chen¹, Ruiming Li¹, Nan Wang¹, Yuankai Luo¹
Rui Wang³, Zipeng Liu¹, Huobin Tan¹, Hongyi Gao^{2*}, Yue Zhang^{3*}, Ge Wang^{2*}

¹School of Computer Science, Beihang University, Beijing & 100191, China.

² School of Materials Science & Engineering, University of Science and Technology Beijing, 100083, China.

³Westlake University, Hangzhou & 310030, China.

*Corresponding author. Email: hygao@ustb.edu.cn, yue.zhang@wias.org.cn, gewang@ustb.edu.cn

†These authors contributed equally to this work.

The extraction of Metal-Organic Frameworks (MOFs) synthesis route from literature has been crucial for the logical MOFs design with desirable functionality. The recent advent of large language models (LLMs) provides disruptively new solution to this long-standing problem. While the latest researches mostly stick to primitive zero-shot LLMs lacking specialized material knowledge, we introduce in this work the few-shot LLM in-context learning paradigm. First, a human-AI interactive data curation approach is proposed to secure high-quality demonstrations. Second, an information retrieval algorithm is applied to pick and quantify few-shot demonstrations for each extraction. Over three datasets randomly sampled from nearly 90,000 well-defined MOFs, we conduct triple evaluations to validate our method. The synthesis extraction, structure inference, and material design performance of the proposed few-shot LLMs all significantly out-

play zero-shot LLM and baseline methods. The lab-synthesized material guided by LLM surpasses 91.1% high-quality MOFs of the same class reported in the literature, on the key physical property of specific surface area.

Introduction

The extraction of material synthesis route from scientific texts using machine learning techniques has long been a popular task in AI for science (1–4) as well as Cheminformatics (5–10). This work explores the topic based on an emerging family of 100k+ crystalline porous materials called Metal-Organic Frameworks (MOFs) (11). MOFs have been widely used in catalysis (12), gas adsorption (13), energy storage (14), and many other fields (15, 16) due to their quantitatively tunable structure and functionality driven by flexible synthesis. These characteristics make MOFs a typical material class where precise and comprehensive knowledge of their synthesis route becomes extremely critical for further material design and application (17, 18).

Currently, there have been 100k+ MOFs successfully synthesized in the laboratory. Their detailed synthesis conditions are basically recorded by academic literature in various textual or tabular formats. Machine learning methods have been applied to the literature text to automatically extract synthesis conditions, ranging from basic pattern recognition methods in the natural language processing (NLP) field (7, 19) to advanced deep learning models (1, 2, 5, 6, 20, 21), and Large Language Models (LLMs) (3, 8, 10, 22). Before the LLM era, the best-performing model using exact-match evaluations (Table 1(A)), i.e. He et al. (5), achieves an F1 score of 0.9, though on a much easier task of precursor name extraction (our model has macro-F1 \geq 0.98 on the same task). In addition, classical supervised learning approaches suffer from complex model building and parameter tuning overhead, and inherent subjective errors due to human-annotated training data (23).

The introduction of LLMs to materials chemistry (22, 24–27) brings a disruptively new solution to the problem of material synthesis route extraction owing to LLMs' strong capability in handling disparate forms of scientific texts with the same off-the-shelf model interface. Among state-of-the-art results (Table 1(B)), the best model for our task is the zero-shot GPT-3.5 by Zheng et al. (8), which achieves a macro-F1 of 0.92 on the SIMM dataset. In this work, by conducting

comprehensive evaluations, we find that the performance of zero-shot LLM falls from 0.92 by subjective evaluation in (8) (or called manual score in (3)) to 0.74 by objective evaluation (or called exact match in (3)), due to its well-known deficiency of lacking specialized knowledge in sparse scenario such as MOFs (28). It is further shown by our computational and material experiments that, the low performance and non-exact-match of zero-shot LLMs let down the material structure inference metrics by 23.9% and key MOFs physical property by 75.1%, when applied to the important tasks of material inference and design.

This paper introduces the few-shot in-context learning paradigm as the standard approach to augment general-purpose LLMs on the material synthesis extraction problem. First, to secure high-quality demonstrations for few-shot LLMs and overcome the accuracy bottleneck of human annotations, we leverage the complementary nature of human expertise and AI intelligence, and propose a human-AI interactive data curation process. Our method enjoys the best of both worlds and offers the highest data quality in ground-truth demonstrations produced. Second, to quantify the optimal number of few-shot demonstrations, we propose to apply the popular retrieval-augmented generation (RAG) algorithms to adaptively select a small number of 4~6 best few-shots combination for each synthesis route extraction. The approach efficiently supports high-throughput synthesis extraction and flexible application deployment by greatly cutting down the input context size for commercial LLMs, in comparison to expensive technologies such as fine-tuning (3, 10) requiring thousands of demonstrations. Finally, the proposed few-shot LLM method is further enhanced with external material knowledge via prompt engineering, and integrated with offline machine learning models and post-processing to form an end-to-end synthesis extraction pipeline for efficiently processing large amount of literature text.

We conduct triple evaluations to validate the proposed method. On three MOFs datasets, the few-shot LLM consistently delivers synthesis extraction accuracy higher than existing methods (0.94 in average macro-F1 vs. 0.77 by zero-shot LLM). To our knowledge, this is the first method that achieves a full-set material synthesis route extraction accuracy above 0.9 under exact-match evaluations. On material structure inference and design evaluations, the proposed few-shot LLM again outperforms state-of-the-art significantly. The actually-synthesized material sample surpasses 91.1% high-quality MOFs reported in the literature, on the BET-measured surface area value.

Table 1: Material info extraction performance reported in the literature where the most relevant task from each paper is listed: (A) classical ML models; (B) LLMs. Here exact match means a true positive is counted only when the extracted info is exactly the same with the annotated ground-truth. Manual/similarity scores mean a true positive can be counted if only the extracted info is considered correct by a human judge (3) or is similar enough, in comparison to the ground-truth.

A	Literature Source	Extraction Task	ML Model	Evaluation Method	Performance Metric
Swain et al. 2016 (19)	Chemical Entity Extraction	Pattern Recognition + CRF	Exact match	F1 = 0.878 (CHEMDNER)	
Kim et al. 2017 (20)	Material Entity Extraction	Neural Network + Pattern Recognition	Exact match	Macro-F1 = 0.81	
He et al. 2020 (5)	Solid-State Synthesis Precursor Extraction	Bi-LSTM	Exact match	F1 = 0.90	
Trewartha et al. 2022 (2)	Material Synthesis Entity Extraction	BERT pre-trained with material knowledge	Exact match	Macro-F1 = 0.75 (average of doping and nanoparticle)	
Gupta et al. 2022 (21)	Material Entity Extraction	SciBERT pre-trained with material literature + BiLSTM + CRF	Exact match	Macro-F1 = 0.781 (average of SOFC, SOFC-Slot, & Matscholar)	
Vaucher et al. 2020 (1)	Material Synthesis Action Sequence Extraction	Transformer	Similarity score	ACC = 0.824 (similarity ≥ 0.75)	
Park et al. 2022 (6)	MOFs Synthesis Route Extraction	Bi-LSTM + CRF + Pattern Recognition	Similarity score	Macro-F1 = 0.903 (similarity ≥ 0.8)	
Glasby et al. 2023 (7)	MOFs Synthesis Route Extraction	Pattern Recognition	Manual score	Macro-F1 = 0.517	

B	Literature Source	Extraction Task	LLM	Learning Method	Evaluation Method	Performance Metric
Zhang et al. 2024 (10)	MOF Synthesis Route Extraction	GPT-3	Fine-tune	Exact match	ACC = 0.827	
Dagdelen et al. 2024 (3)	MOFs Info Extraction	GPT-3	Fine-tune	Exact match	Macro-F1 = 0.519	
Polak et al. 2024 (22)	Material Single Property Extraction	GPT-4	Zero-shot with Reflection	Between exact match and manual score	Between exact match and manual score F1 = 0.892 (constrained data), F1 = 0.873 (practical data)	
Zheng et al. 2023 (8)	MOFs Synthesis Route Extraction	GPT-3.5	Zero-shot	Manual score	Macro-F1 = 0.92	

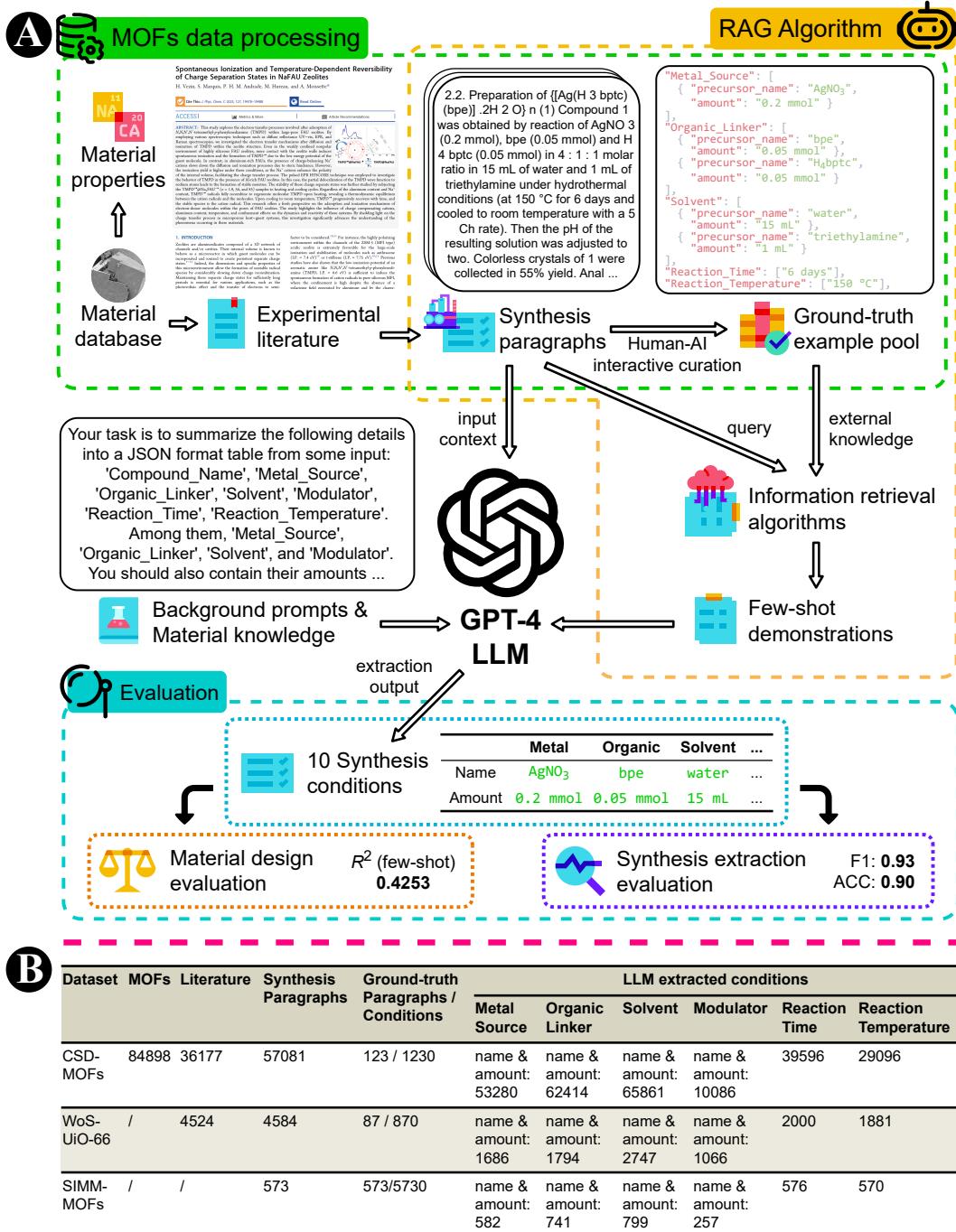


Figure 1: Overview of our MOFs synthesis route extraction proposal using few-shot LLM in-context learning method. (A) The technical workflow is composed of multiple core components including data processing on material database, the RAG algorithm to select few-shot demonstrations, the LLM engine, and the multi-faceted material evaluation; (B) Statistics of MOFs datasets used in this work.

Results

Our technical workflow is given in Fig. 1(A). The MOFs experimental literature is collected from three data sources (Fig. 1(B), (Supplementary Text)). The first is the CSD material database v5.43 (29) where we download 36,177 papers covering 84,898 unique MOFs (CSD-MOFs), the second is Web of Science platform (30) where we retrieve all the 4,524 papers related to the class of UiO-66 MOF with Zr as metal (WoS-UiO-66), and the last is the SIMM data from Zhang et al. (10) containing 573 MOFs' synthesis route manually annotated over Zheng et al. (8)'s raw data. The full-text of each paper is pre-processed to locate paragraphs relevant to MOFs synthesis (Materials and Methods). The GPT-4 LLM (31) is employed to extract 10 essential conditions from each paragraph (Table 3). The synthesis extraction result is first evaluated on their literal accuracy with respect to ground-truth data, and then tested on the real-world scenarios of MOFs structure inference and synthesis. Take the CSD-MOFs dataset for example, the proposed few-shot LLM method achieves much higher extraction accuracy in 7 out of 10 synthesis conditions (Fig. 2(A), macro-F1=0.93), than the baseline zero-shot LLM (Fig. 2(B), macro-F1=0.81). This advantage is consistent on all 8 state-of-the-art LLMs tested (Fig. 2(C) vs. (D)). The performance at the other two datasets also show similar comparison results ((Supplementary Text), Fig. S3, Fig. S5). Our fully-tuned high-throughput synthesis extraction workflow now processes over 500 million of scientific texts from all available MOFs literature within 7 hours.

Human-AI Interactive Data Curation

A prerequisite to our few-shot LLM method is to obtain a high-quality example pool, which should contain the ground-truth synthesis conditions annotated over a set of sample synthesis paragraphs (see an example in the RAG block of Fig. 1(A)). Otherwise, the few-shot examples without ground-truth annotation will lead to worse performance even than the zero-shot LLM without examples (red+square lines in Fig. 3(B)). We first apply a standard human annotation protocol (Supplementary Text) and software (Fig. S2). On the CSD-MOFs dataset, the few-shot LLM using human-annotated example pool achieves a best synthesis extraction macro-F1 score of 0.87 (orange+triangle lines in Fig. 3(B)), lagging behind our F1 target of 0.9.

We then consider the latest approach of automated AI annotation, in which LLM is initially

	4-shot BM25 by GPT-4 Turbo												4-shot BM25 by Different LLMs								
	F1	0.99	0.97	0.87	0.87	0.98	0.97	0.87	0.87	0.93	0.93	0.93	F1	0.93	0.91	0.88	0.91	0.90	0.92	0.91	0.88
ACC	0.99	0.95	0.77	0.92	0.96	0.94	0.76	0.92	0.89	0.91	0.90		ACC	0.90	0.89	0.83	0.88	0.87	0.89	0.88	0.85
Rec.	1.00	0.98	0.99	0.92	1.00	0.98	0.99	0.92	0.95	0.95	0.97		Rec.	0.97	0.97	0.97	0.97	0.96	0.97	0.98	0.97
Pre.	0.99	0.96	0.78	0.83	0.97	0.95	0.77	0.83	0.90	0.91	0.89		Pre.	0.89	0.86	0.80	0.86	0.86	0.87	0.86	0.82
metal name													GPT-4 Turbo								
organic name													GPT-40								
solvent name													GPT-3.5								
modulator name													Claude 3.5 Sonnet								
metal amount													Claude 3.5 Haiku								
organic amount													Gemini-1.5 Flash								
solvent amount													Gemini-1.5 Pro								
modulator amount													GLM-4-plus								
reaction time													average								

	Zero-shot by GPT-4 Turbo												Zero-shot by Different LLMs								
	F1	0.96	0.95	0.74	0.73	0.83	0.82	0.74	0.63	0.82	0.91	0.81	F1	0.81	0.78	0.72	0.69	0.51	0.81	0.68	0.56
ACC	0.93	0.90	0.59	0.85	0.71	0.70	0.58	0.80	0.72	0.88	0.77		ACC	0.77	0.74	0.71	0.66	0.47	0.77	0.65	0.50
Rec.	0.97	0.96	0.95	0.74	0.96	0.95	0.95	0.69	0.97	1.00	0.91		Rec.	0.91	0.95	0.85	0.96	0.39	0.96	0.94	0.46
Pre.	0.96	0.93	0.61	0.72	0.73	0.72	0.60	0.58	0.70	0.83	0.74		Pre.	0.74	0.68	0.66	0.59	0.75	0.72	0.58	0.72
metal name													GPT-4 Turbo								
organic name													GPT-40								
solvent name													GPT-3.5								
modulator name													Claude 3.5 Sonnet								
metal amount													Claude 3.5 Haiku								
organic amount													Gemini-1.5 Flash								
solvent amount													Gemini-1.5 Pro								
modulator amount													GLM-4-plus								
reaction time													average								

Ground-truth												
LLM Output	TP: LLM output (not empty) = ground truth (not empty)						FP: LLM output (not empty) ≠ ground truth					
	FN: LLM output (empty) ≠ ground truth (not empty)						TN: LLM output (empty) = ground truth (empty)					

Figure 2: Synthesis route extraction performance on the CSD-MOFs dataset. Key indicators (F1, ACC, Recall, Precision) are listed. Only literature with annotated ground-truth are included. (A) the proposed few-shot LLM (GPT-4 Turbo) with RAG algorithm; (B) zero-shot LLM as the baseline; (C) few-shot approach on 8 different LLMs; (D) zero-shot approach on 8 different LLMs; (E) confusion matrix definition for performance evaluation.

applied in a zero-shot mode to extract all synthesis conditions. The zero-shot LLM output is then used as ground-truth annotations in a 2nd-round few-shot LLM in-context learning, which generates the final AI annotations. It is shown through experiments that the best few-shot LLM using

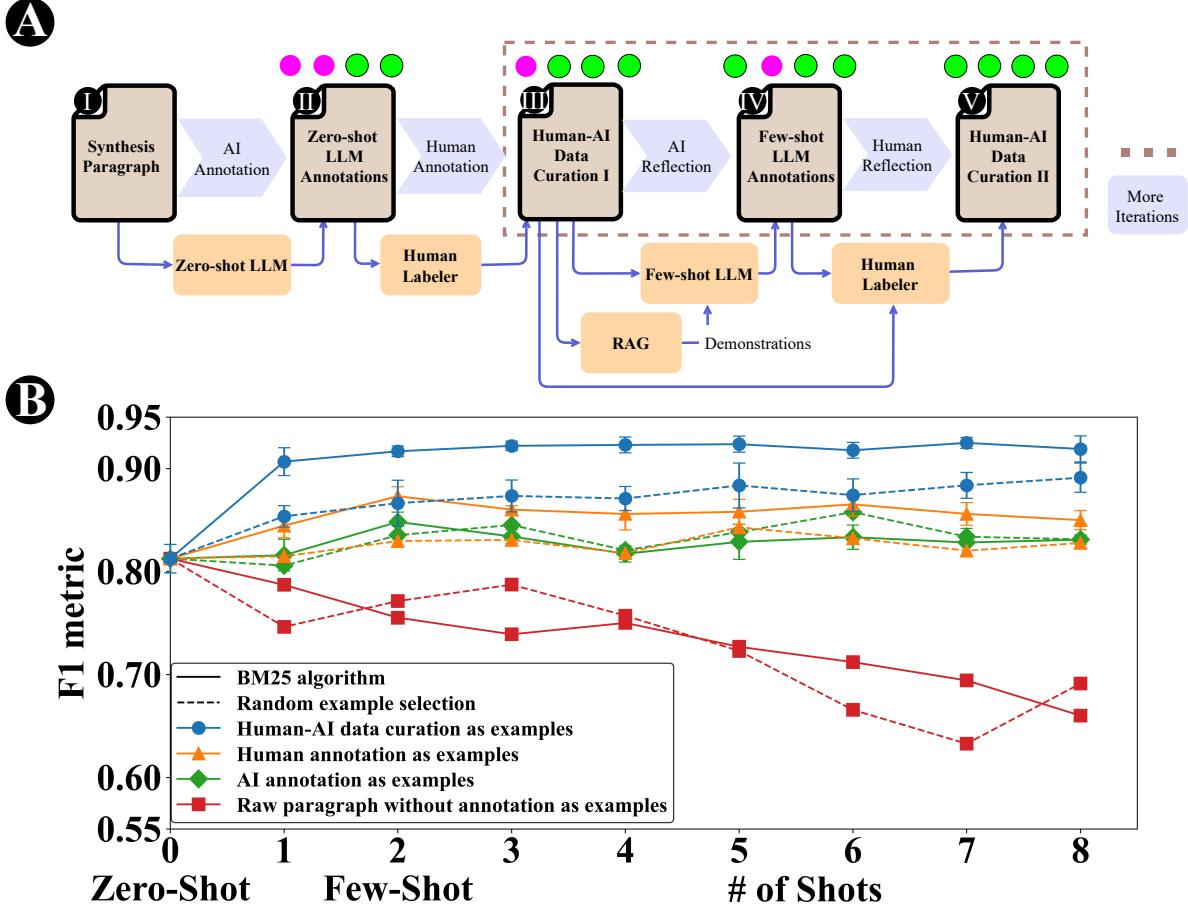


Figure 3: Human-AI interactive data curation method. (A) the overall workflow to improve the data quality of few-shot examples; (B) the synthesis extraction performance by different data curation methods and varying number of shots. The 95% CI error bar by 5 repeated tests is only displayed on the human-AI curation plot to reduce visual clutter.

AI annotated example pool also achieves a macro-F1 of 0.86 (green+diamond lines in Fig. 3(B)). Neither human annotations nor purely AI-generated examples achieve the best data quality. In fact, human expertise and AI’s capacity in labeling ground-truth synthesis conditions are complementary to each other (more in (Discussion)).

We propose a new approach of human-AI interactive data curation to retrieve the best-quality ground-truth examples. As shown in Fig. 3(A).I, raw synthesis paragraphs are first processed by LLM in a zero-shot mode to obtain an initial AI annotation (Fig. 3(A).II). Human labelers then work on the AI annotation and achieve a best-effort human annotation (Fig. 3(A).III), which is the first round of human-AI interactive data curation. The complementary advantage here lies in that

AI generates annotation in a highly efficient way and greatly alleviates the fatigue issue of human labelers, who can focus on rare cases requiring specialized knowledge. In the second round, the few-shot LLM is applied with the human annotation output in the first round as ground-truth examples. The output (Fig. 3(A).IV) represents a reflection by AI that resolves potential random errors made by human labelers in the last round. Finally, human labelers combine the latest human and AI annotations, and achieve the second round of human-AI interactive data curation (Fig. 3(A).IV). Here only the cases where human and AI disagree with each other are re-examined. In fact, this human-AI interactive curation process, as in the dashed frame of Fig. 3(A), can iterate more than one round. In our work after the second reflection, an excellent performance has already been achieved (macro-F1=0.93, blue+circle lines in Fig. 3(B)). The final ground-truth example pool on CSD-MOFs dataset includes 123 rows of 1230 synthesis conditions in total.

Few-Shot Large Language Model with Material Knowledge

At the core of our technical workflow (Fig. 1(A)), we apply the GPT-4 Turbo LLM, which exhibits the highest extraction performance among state-of-the-art LLMs (Fig. 2(C)). The latest few-shot in-context learning approach (FS-ICL (28)) is introduced, which refers to a typical learning paradigm to adapt the task-agnostic language models to various downstream tasks while achieving optimized performance on each task. In more detail, FS-ICL takes a few prompted examples as input (known as shots), each composed of a context and a labeled completion, in addition to background prompts such as task description. For the task of MOFs synthesis route extraction, a context refers to a paragraph containing all the synthesis conditions of a MOF and the labeled completion refers to the ground-truth synthesis conditions annotated and curated by human and AI in our work. The final LLM extraction is made by prompting a new context and asking the language model to complete it.

The FS-ICL approach enjoys high flexibility to work on many tasks (e.g., synthesis extraction of various materials) without the need to re-train the model in contrast to fine-tuning (32, 33). Yet, FS-ICL still faces multiple challenges. First, the cost induced by few-shots to LLM needs to be quantified. Second, the prompt format in FS-ICT (e.g., the wording and ordering of examples) can bring uncertainty to the extraction performance. How to alleviate this uncertainty remains unknown.

Using RAG to optimize few-shot data quality and quantity

We introduce the RAG algorithm (34, 35) which retrieves the best K -shot examples for each

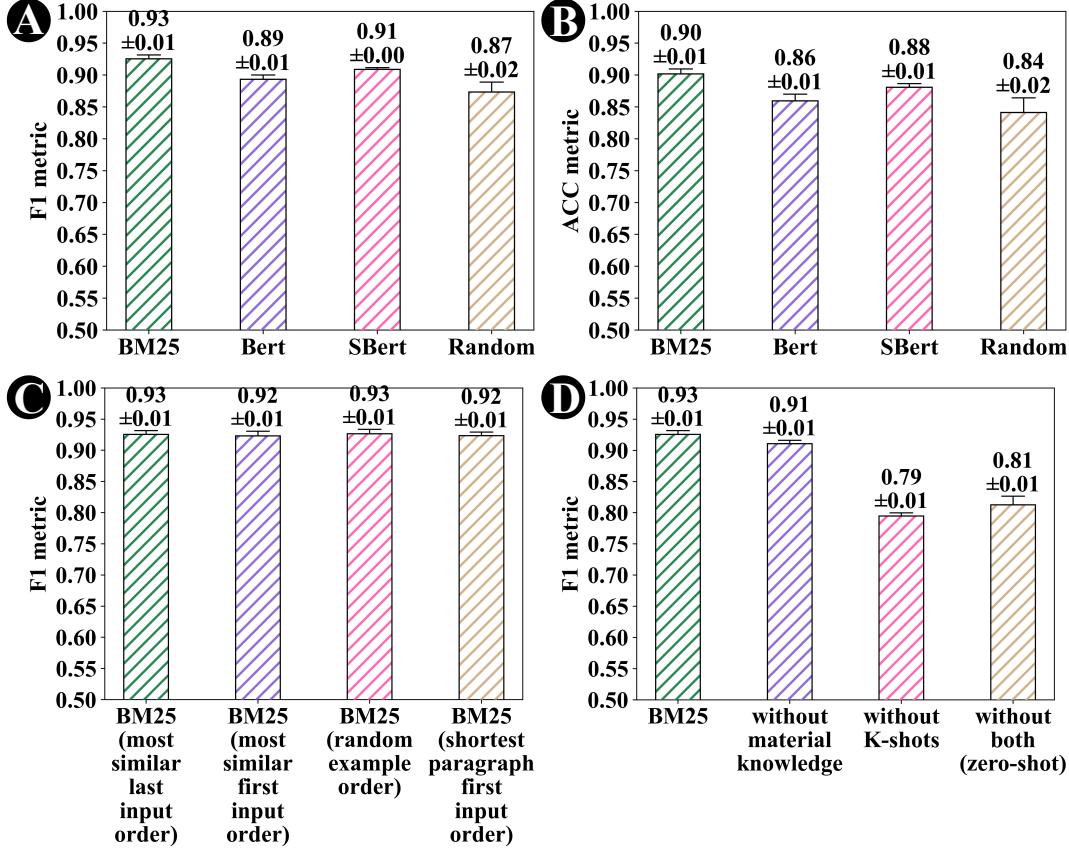


Figure 4: The synthesis route extraction performance by RAG configurations. (A)(B) comparison of RAG algorithms, on F1 and ACC respectively; (C) comparison of different few-shot orders; (D) the impact of LLM prompt compositions. Error bars indicate 95% CI within 5 repeated tests.

input context to augment the LLM and then generates the predicted completion. Three mainstream RAG algorithms are applied here: BM25 (36), BERT (37), and Sentence-BERT (SBERT) (38). They differ in how to compute the similarity between the input context and candidate examples (Materials and Methods). We then conduct an experiment to compare these RAG algorithms, using the CSD-MOFs dataset. By default, a typical setting of $K = 4$ is used.

Experimental results in Fig. 4(A)(B) indicate that the BM25 algorithm (first column) achieves the best synthesis extraction performance among all the compared algorithms, with a macro-F1 of 0.93 and overall accuracy (ACC) of 0.90. The result is quite stable for each algorithm within 5 repeated runs, as shown by the 95% CI error bar (0.006 for BM25 on F1). Notably, any of the tested algorithms is significantly better than a random selection of examples on F1 (the last column), e.g., $t(8) = 4.68$, $p = .0016$ by a two-tailed t-test comparing Bert and random algorithms.

This showcases the effectiveness of RAG mechanism. On the best BM25 algorithm, we further test the impact of few shots' input order within the LLM prompt. As shown in Fig. 4(C), the differences are not significant among all the tested ordering strategies ($p > 0.05$ by two-tailed t-tests). We decide to fix the few-shot input order to the most similar last setting with the best performance. Next, we evaluate the impact of composition of LLM prompts. The ablation study results in Fig. 4(D) demonstrate that the few-shot examples are the primary drive of performance improvement, followed by the material knowledge provided as background prompts. We then study the optimal few-shot quantity required by RAG algorithms (i.e., K). Fig. 3(B) and Fig. S1 illustrate that, with the best BM25 RAG algorithm and the proposed human-AI ground-truth annotation (solid blue line with circle symbol), both F1 and ACC increase the most from zero-shot to one-shot, and continue to grow until the peak of $K = 4$ ($F1 = 0.93$, $ACC = 0.90$). Meanwhile, the few-shot method with random example selection (dashed blue lines) shows consistently lower performance than the BM25 algorithm. This result again demonstrates the effectiveness of the proposed RAG algorithm.

Sizing the few-shot example pool

After optimizing both the quantity and quality of few-shot examples, we then study how large an example pool with ground-truth annotations is required for high-throughput MOFs synthesis extraction. Another experiment is set up that assumes the full dataset to be the 123 synthesis paragraphs of CSD-MOFs with known ground-truth conditions. The example pool is randomly chosen from the dataset. In another setting, a subset of 60 ground-truth synthesis paragraphs of CSD-MOFs is used as the full dataset.

Fig. 5(A) illustrates the impact of example pool size on the MOFs synthesis extraction performance. The initial increase of example pool size (from 0 to 5 in the figure) contributes the most performance gain, regardless of the size of the full dataset. This is coherent with the effect observed on zero-shot vs. one-shot learning. More annotations and larger example pools will almost always bring performance gains and less uncertainty, up to 66.7% and 52.8% of the full dataset, in the two settings respectively. A smaller dataset (green+triangle line) requires a pool with fewer examples than a larger dataset (red+circle line), in achieving the same level of performance.

Fig. 5(B) further demonstrates the impact of both example pool size and K -shots. With an example pool size no smaller than 45, the extraction performance becomes indistinguishable with

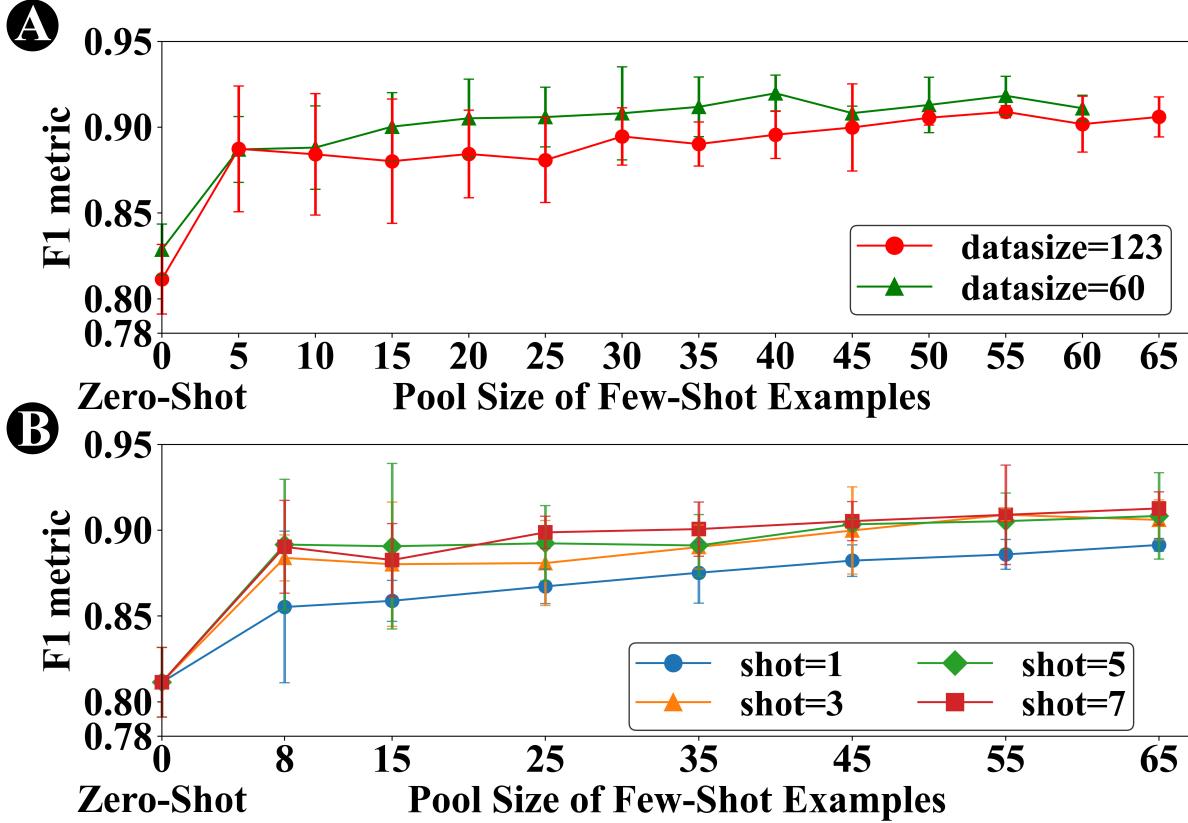


Figure 5: MOFs synthesis route extraction performance using example pool of varying sizes.

(A) average F1 and its 95% CI on the 123-paragraph and 60-paragraph CSD-MOFs datasets ($K=3$ for both); (B) on the 123-paragraph dataset using different K -shots.

shot settings of $K \geq 3$. Meanwhile, the one-shot learning has constant a performance gap, suggesting a setting of $K \geq 3$ at least.

MOFs Structure Inference

We evaluate the few-shot synthesis extraction method in a real-world MOFs synthesis-structure inference task. The task is to predict the microstructure properties of MOFs, namely global cavity diameter, pore limiting diameter, largest cavity diameter, and framework density, using the extracted synthesis conditions including metals, organic linkers, solvents, modulators, and reaction duration/temperature. The few-shot/zero-shot LLMs and another classical machine learning (ML) algorithm proposed by Park et al. (6) are compared. Each method is used to extract 10 synthesis conditions from each unique paragraph of the 5269 MOFs in the CSD-MOFs dataset. The raw textual conditions extracted are post-processed by methods in (Materials and Methods) into densely

Table 2: Performance comparison of the MOFs framework density inference task using different machine learning models. Left: few-shot vs. zero-shot LLMs on the CSD-MOFs dataset; Right: few-shot LLMs vs. classical machine learning algorithm on the KAIST dataset (6).

Inference Model	CSD-MOFs dataset				KAIST dataset			
	Zero-shot		Few-shot		Classical ML		Few-shot	
	R ²	RMSE	R ²	RMSE	R ²	RMSE	R ²	RMSE
KNN	0.2109	0.4440	0.2599	0.4315	0.1841	0.5310	0.2058	0.5239
Bayesian Ridge Regression	0.2085	0.4448	0.2871	0.4222	0.2110	0.5222	0.2215	0.5187
Lasso Regression	0.2124	0.4436	0.2958	0.4198	0.2079	0.5236	0.2190	0.5187
Neural Network	0.2411	0.4392	0.3121	0.4172	0.2130	0.5214	0.2208	0.5201
Random Forest	0.2974	0.4193	0.3976	0.3880	0.2633	0.5047	0.2907	0.4947
XGBoost	0.3403	0.4061	0.4253	0.3790	0.2644	0.5042	0.3097	0.4884

distributed vector representations.

Six ML models are applied for the structure inference: KNN, Lasso Regression, Bayesian Ridge Regression, Neural Networks, Random Forest, and eXtreme Gradient Boosting (XGBoost). The performance is evaluated by the test-set coefficient of determination (R^2) in a 10-fold cross-validation. On four microstructure properties inferred, the first three properties irrelevant to MOFs synthesis conditions lead to negative or close to zero R^2 in all models, regardless of the extraction method. The last property of MOFs framework density is mostly predictable with synthesis conditions, with the best R^2 value larger than 0.4 (Table 2). It can be observed in Table 2 that the synthesis conditions obtained by the few-shot LLM enjoy much larger predictive power than those obtained by the zero-shot LLM, on all six ML models. The R^2 values are 31.4% higher on average and the difference is statistically significant under a paired two-tailed t-test ($t(5) = 11.15, p = .0001$). Between the few-shot LLM and the classical ML algorithm, as shown in Table 2, the few-shot LLM achieves 8.9% higher R^2 than the classical ML in average with all the 6 predictive models. The difference is statistically significant under a paired t-test ($t(5) = 3.55, p = .016$). As the MOFs data size in this trial is smaller than the few-shot vs. zero-shot experiment (Supplementary Text), the absolute R^2 values are mildly smaller. The comparison on the Root Mean Squared Error (RMSE)

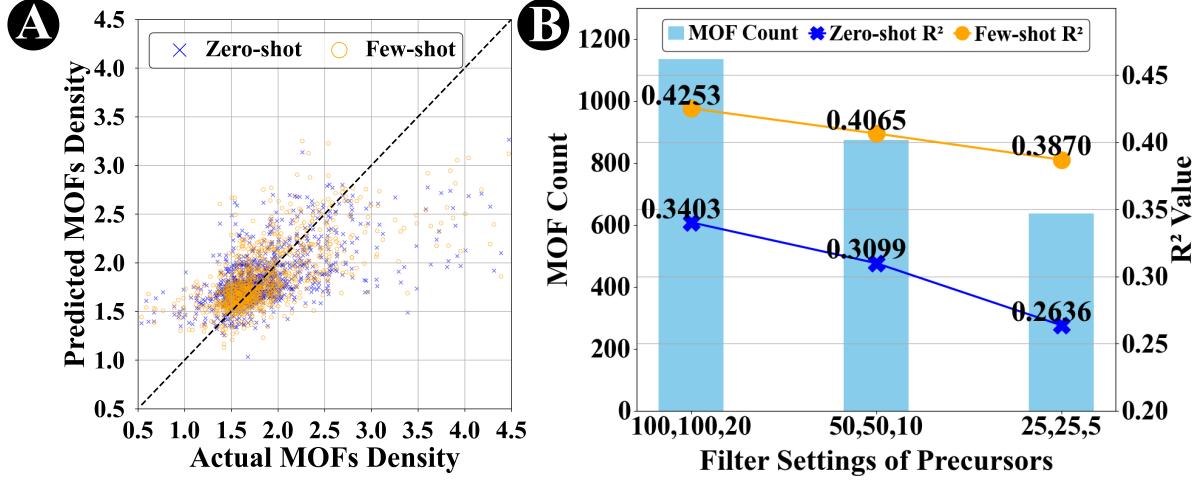


Figure 6: Detailed MOFs framework density inference performance. (A) predictive power of the best XGBoost model, few-shot LLM vs. zero-shot LLM; (B) applying different precursor filters.

metric (Table 2) validates the same result that the few-shot LLM enjoys lower predictive error than the zero-shot method and classical ML.

Drilling down to more details, we illustrate the best XGBoost inference result on the scatterplot of Fig. 6(A). It reveals that the actual vs. predicted MOFs density distribution of the few-shot LLM (orange circles) exhibits higher proximity to the optimal prediction line (black dashed line), than the predictions by zero-shot LLM (blue crosses). We also gradually reduce the evaluation dataset by enforcing stricter precursor filters and selecting only higher-ranked synthesis condition values. As shown in Fig. 6(B), the R^2 of predictive models by few-shot LLM drops mildly, much slower than the decrease of zero-shot LLM. The result demonstrates that the proposed few-shot LLM not only extracts more accurate synthesis conditions, but also significantly improves the downstream material inference task.

Real-World Synthesis of New MOFs

The synthesis conditions extracted by our method and the baseline are compared in suggesting the best synthesis routes for designing pure UiO-66 (Zr) MOF with the largest specific surface area (SA). The class of UiO-66 MOFs is selected because it is popular in the research community and many literature is available for its synthesis. The mutable synthesis conditions of UiO-66 (Zr) are also limited, with metal and organic precursors fixed, so that most of the suggested synthesis routes

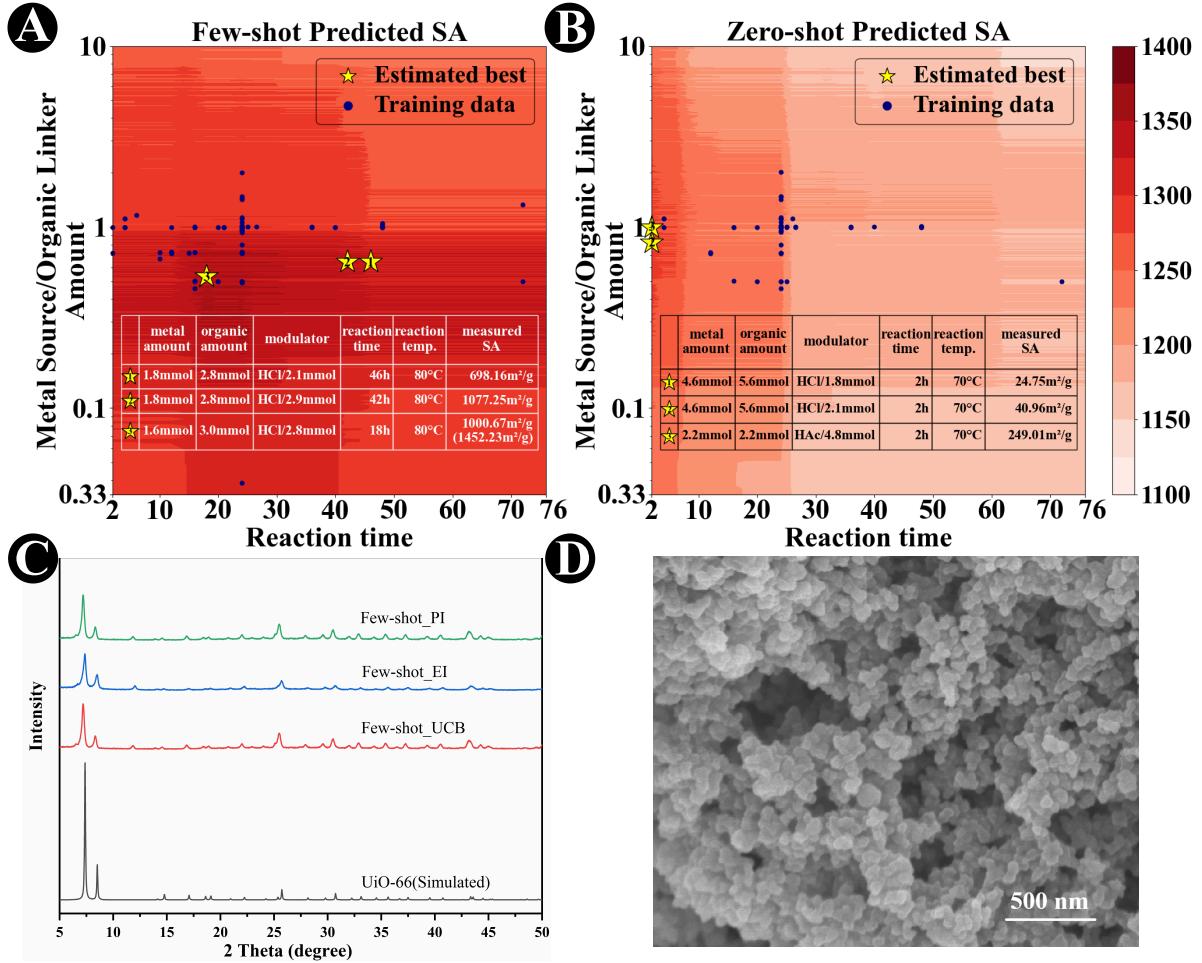


Figure 7: Real-world MOFs design assisted by LLM. (A) few-shot LLM extracted synthesis conditions (blue points), the SA inference result (heat map), and the suggested optimal conditions for the largest SA value (yellow stars); (B) results by zero-shot LLM for comparison; (C) XRD interpretation of lab-synthesized UiO-66 MOFs suggested by few-shot LLM; (D) SEM result on one of the MOFs suggested by few-shot LLM.

are feasible for comparison. The experiment first builds a SA inference model by the random forest algorithm, using the extracted synthesis routes on 261 pure UiO-66 MOFs from the WoS-UiO-66 dataset (Supplementary Text).

Fig. 7(A)(B) present the SA inference model using few-shot and zero-shot LLM-extracted training data (blue points in the figure), respectively. The two most influential synthesis conditions, i.e., reaction time and the ratio of metal precursor amount to organic linker amount, are used as projection dimensions. The heat map in the figures illustrates the maximal SA value inferred

from the corresponding synthesis conditions projected to the same point in the view. We apply the Bayesian optimization algorithm (Materials and Methods) to suggest three best synthesis routes for the maximally expected SA value of UiO-66, as indicated by the yellow stars in each figure. The six lab-synthesized UiO-66 samples suggested by few-shot and zero-shot LLMs are gone through BET test to measure their actual SA values. As shown in the embedded table of Fig. 7(A)(B), the measured SAs of MOFs by few-shot LLM are more than four times as large as the corresponding SA suggested by zero-shot LLM. The SA inference by zero-shot LLM shows a biased model preferring smaller reaction time in UiO-66 MOF synthesis, as in Fig. 7(B), which can explain its bad performance in MOFs design. UiO-66 (Zr) MOFs have slow reaction kinetics, rendering the suggested 2-hour reaction time insufficient to achieve a well-formed crystalline structure. In comparison, the synthesis output suggested by the few-shot LLM is proven to be pure UiO-66, as shown by the XRD pattern in Fig. 7(C) w.r.t. the simulated UiO-66. The SEM image in Fig. 7(D) also shows very high crystallinity in this production. Note that the MOFs measured in the literature mostly go through advanced rinsing and activation processes to increase SA values. We also process one of our best few-shot LLM produced MOF samples with standard activation. The SA value increases from 1000.67 to 1452.23, outperforming 91.1% UiO-66 MOFs' SA in our literature collection. This showcases the great potential to apply the LLM synthesis extraction method to guide high-performance material design.

Discussion

This work studies the new paradigm of applying few-shot LLM in-context learning to the important problem of MOFs synthesis route extraction. It is shown through experiments that both the quality and the quantity of few-shot demonstrations are critical to the LLM performance on MOFs synthesis route extraction. We introduce a novel approach of human-AI interactive data curation to enhance few-shot demonstration quality and a calibrated BM25 RAG algorithm to size the optimal few-shot quantity. Experimental result reveals that a small overhead of 4~6 shots already achieve excellent performance for few-shot LLM on three typical MOFs datasets (average macro-F1 = 0.94), which are significantly better than the state-of-the-art approach by zero-shot LLM (average macro-F1 = 0.77). Practical issues regarding high-throughput MOFs synthesis extraction and downstream applications

are resolved using elaborate methods including offline synthesis paragraph detection, data post-processing, and LLM prompt engineering. Our proposal is also thoroughly evaluated over real-life MOFs structure inference and high-performance design tasks. Compared with the baseline zero-shot LLM, the few-shot LLM method increases the MOFs framework density inference performance by 31.4% and improves the surface area of lab-synthesized UiO-66 MOFs by more than three times. The lab-synthesized material guided by LLM surpasses 91.1% high-quality MOFs of the same class reported in the literature, on the key physical property of specific surface area.

In addition, we summarize the following implications by this work.

Superiority of Human-AI Interactive Data Curation

We consulted MOF experts to evaluate all errors produced by the few-shot LLM method when human-only annotations are used as ground-truth demonstrations. Out of 261 reported errors, 103 LLM outputs (39.5%) were identified as actually correct by the expert's reflection, 38 (14.6%) had certain issues but contributed to refining the corresponding ground-truth, and only 120 (45.9%) were true errors. In 54.1% of cases when human and AI extraction have conflicts, AI can help to consolidate a better ground-truth example. This implies that human and AI can be complementary on the task of labeling ground-truth material synthesis conditions. The joint human-AI approach then enjoys three superiorities. First, though human labelers are excellent in the usage of material knowledge, they can fail to strictly follow pre-defined annotation rules. For example, to standardize the solvent condition, it is required to leave out all modifiers of a common solvent. Human annotators often extract “hot water” instead of “water”, as s/he mostly focuses on knowledge extraction but neglects obligatory rules. Humans are poor multi-objective task executors compared with AI. AI introduces less errors when rules are provided in either background prompt or examples. Second, human labelers can suffer from fatigue when working with a large number of annotation tasks, leading to random errors, e.g., missing or adding a few characters/words. AI can be applied to eliminate this issue: a zero-shot LLM annotation in the first round reduces human efforts, their fatigue, and the resulting random errors. Third, general-purpose LLMs alone lack long-tailed material knowledge not appearing in both background prompts and demonstrations. For example, LLM can fail to correctly extract a minor MOF synthesis route appearing very few times in the literature. Human can fix such labeling errors using his/her strong generalization capability.

Potential Extension to New Materials and Applications

The cost to apply our method to extract a new material’s synthesis conditions includes: first, the set of experimental material literature should be supplied. This can be off-the-shelf from the users or acquired from academic databases such as Web of Sciences (30). Second, human input is necessary which specifies the exact knowledge to be extracted and its example form in the literature as the seed for few-shot demonstrations. Our method provides an end-to-end solution to augment the human-specified knowledge into background prompts and few-shot demonstrations. More importantly, the number of required annotations, i.e., the example pool size, can be estimated by the empirical result of our work. Third, optionally the input prompt can be tuned per the characteristics of each material to improve the performance. Finally, by introducing newer versions of LLMs such as GPT-4v, extracting from not only textual forms but also pictures and videos can be envisioned.

Materials and Methods

Synthesis Paragraph Detection

By the latest GPT-4 turbo pricing (\$10 per 1M tokens), a single pass of LLM over all the 100k MOFs literature can sum up to a non-negligible cost of \$10k (assuming 10k words per literature), while performance tuning normally requires several passes. After investigating a large set of MOFs synthesis literature, it is found that in most cases the authors state the detailed synthesis route of a MOF only in a few paragraphs of the paper, e.g., those starting with “Synthesis of [a MOF’s chemical formula]”. By estimation, the length of these synthesis paragraphs (est. 600 words) only occupies 6% of the content of a paper in average. By applying an offline synthesis paragraph detection model, the financial cost in using commercial LLMs is reduced by 94% and the synthesis extraction performance and efficiency are also improved as LLM can now focus on a much smaller set of textual content for extraction.

To this end, we build a binary classification model to determine whether a paragraph contains a suite of synthesis route. 440 papers from the CSD-MOFs dataset and 87 papers from the WoS-UoI-66 dataset are sampled as the training dataset, where ground-truth synthesis paragraphs are manually annotated as positive samples (Supplementary Text). The remaining paragraphs are used as negative samples. In total, 1349/11783 and 87/852 positive/negative synthesis paragraphs are

obtained on the two datasets. A standard BERT model (39) is applied for the classification. A stratified 5-fold cross-validation is used to evaluate on the imbalanced data. Our model achieves F1 = 0.951, ACC=0.989 (CSD-MOFs dataset), and F1 = 0.956, ACC=0.979 (WoS-UoI-66 dataset), on detecting (positive) synthesis paragraphs. Finally, we detect 57081 synthesis paragraphs from 36177 papers on the CSD-MOFs dataset, and 4584 synthesis paragraphs from 4524 papers on the WoS-UoI-66 dataset.

Few-shot RAG Algorithms

We apply few-shot RAG algorithms (34) (35) to improve the LLM performance, which retrieves K demonstrations for each synthesis extraction from a pool constructed by human-AI interactive data curation. Formally, given a pool of demonstrations $D = \{d_1, d_2, \dots, d_n\}$ and an input paragraph p , the RAG algorithm retrieves top K similar demonstrations by:

$$K\text{-shots} = \text{sort}((\text{score}(p, d_i), d_i)_{i=1}^n)[:, K] \quad (1)$$

Here, the score function is used to estimate the similarity between document d_i and paragraph p . The function can be categorized into two classes by the document representation method: sparse vector encoders (e.g., TF-IDF, BM25 (36)) and semantic dense vector encoders (40) (e.g., SBERT (38), BERT (37)). We select the best-performing RAG algorithm, i.e., BM25, as the final choice, according to experimental results. BM25 is a probabilistic information retrieval model that ranks documents based on the frequency of query terms within the documents. It balances term frequency (how often a term appears in a document) with inverse document frequency (how rare a term is across the entire document set), thus giving more weight to terms that are significant. The scoring function of BM25 between a paragraph with n terms and a document in a pool of length N is defined as:

$$\text{Score}(p, d) = \sum_{i=1}^n \text{IDF}(p_i) \cdot \frac{f(p_i, d) \cdot (k_1 + 1)}{f(p_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\text{avg_dl}})} \quad (2)$$

$$\text{avg_dl} = \frac{1}{N} \sum_{j=1}^N |d_j| \quad (3)$$

where $f(p_i, d)$ is the term frequency of p_i in document d , $|d|$ is the length of document d , avg_dl is the average length of all documents in the demonstration pool, $\text{IDF}(p_i)$ is the inverse document

frequency of term p_i . k_1 and b are hyperparameters of the model and we use the default BM25 settings of $k_1 = 1.5$ and $b = 0.75$ in our experiment.

For the alternative class of semantic information retrieval, we utilize the document embedding by pre-trained language models to compute the similarity:

$$\text{Score}(p, d) = \frac{f(p) \cdot f(d)}{|f(p)||f(d)|} \quad (4)$$

where $f(x) = \text{PLM}(x)$ denotes the document embedding and PLM refers to a pre-trained language model such as SBERT. The document embedding can be derived by either averaging the token embeddings (mean pooling) or using the [CLS] token embedding (37).

Note that when evaluating RAG algorithms, we mainly use the F1 and ACC of LLM extraction result following the standard definition. The basic metrics of TP (true positive), FP (false positive), TN (true negative), FN (false negative) are first computed. The LLM extraction of each synthesis condition will be classified into one of TP/FP/TN/FN by comparing with the predefined ground-truth annotation, as described in the confusion matrix of Fig. 2(E).

LLM Prompt Engineering for Material Knowledge Augmentation

In addition to few-shot demonstrations, another way to augment the domain knowledge of general-purpose LLM is through the background prompt (41). The previous LLM adaptation on MOFs synthesis extraction by Zheng et al. (8) introduces a preliminary prompt engineering approach, which includes the task description of MOFs synthesis extraction and the output format specification. In our work, based on the latest prompt engineering catalog (42), we propose to further incorporate two types of material knowledge into the background prompt (Fig. S18): definition of each MOF synthesis condition, and deterministic constraints on each condition's numerical/textual value or structure (if any). As shown in Fig. 4(D), by augmenting material knowledge, the macro-F1 metric increases from 0.91 to 0.93. However, when the few-shot examples are not incorporated, the background material knowledge will not lead to improvement by itself.

The list of newly introduced MOFs synthesis definitions and constraints as background prompts is provided in Table 3. We summarize three types of constraints on synthesis conditions: *numerical* that the value of a condition should fall into certain range according to prior knowledge, *textual* that an extracted textual condition should adhere to certain format to speedup follow-up material

Table 3: Material knowledge incorporated as part of LLM background prompts. Both the definition of 10 MOFs synthesis conditions and their numerical/textual/structural constraints are included.

Conditions	Definition	Constraints by Type
Metal Precursor (name & amount)	A metal precursor compound containing metal ions that form the MOF structure	<i>Textual:</i> Only include adjectives modifying the metal precursor itself. Exclude any adjectives...
Organic Linker (name & amount)	The organic precursor linking metal ions or clusters in the MOF	N/A
Solvent (name & amount)	The liquid medium in which reactants are dissolved	<i>Textual:</i> If a solvent contains water, the solvent should include "solution" because...
Modulator (name & amount)	The modulator aims to adjust the reaction condition, such as pH value	<i>Structural:</i> The elements of modulator will not become the backbone of MOF structure after synthesis...
Reaction Process (duration & temperature)	The synthesis process that produces the MOF materials	<i>Numerical:</i> The reaction duration will last several minutes to hours... <i>Structural:</i> Crystallization is not a reaction process...

application, and *structural* that certain rules related to the condition are followed in all MOFs synthesis process. The full background prompt is provided in (Supplementary Text).

Post-processing of Extracted Synthesis Routes

The LLM-based method extracts the synthesis routes precisely as they appear in the literature. However, the same synthesis condition can have different writing styles. Post-processings can help to standardize these conditions so that the downstream material inference from synthesis route will have denser input data to achieve better performance.

Coreference Resolution

MOFs literature often use proxy words like “L” or “H2L” to represent specific organic linkers, a prototype called coreference in NLP. These proxy words are often defined before the synthesis paragraph in the same literature. Due to different writing styles, regular expression based extraction is not enough for resolving all the coreference of proxy words. We introduce a hybrid method combining LLM and regular expression: first, the synthesis paragraph is located in the literature and all the text before the paragraph is input to LLM. The LLM is asked to extract all anaphoric references and their original words. Second, a regular expression is designed to identify coreference proxy words from all the LLM-extracted organic linker conditions. Finally, the proxy words in the extracted synthesis route are matched with the detected anaphoric references. For each match, the proxy word is resolved into its original word. In all the 5269 synthesis paragraphs of the CSD-MOFs dataset, we detect 578 coreferences on the organic linker condition. 79% of them can be resolved by our method. Only 2.3% synthesis paragraphs still have unresolved proxy words in their organic linker condition.

Data Cleansing of Textual Conditions

Textual synthesis conditions, such as metal, organic linker, solvent, and modulator names, often have different representations for the same substance, e.g., both “ H_2O ” and “Water” represent water. We propose a three-step post-processing approach to merge them together.

First, by similarity-based disambiguation, we generate a list of assimilated names to merge based on the similarity score between any two names s_i and s_j :

$$\text{Similarity} = 1 - \frac{L(s_i, s_j)}{\max(|s_i|, |s_j|)} \quad (5)$$

where $L(\cdot, \cdot)$ denotes the Levenshtein distance and $|\cdot|$ denotes the length of a name. By default, we use a similarity threshold of 0.9.

Second, we apply GPT-4 to conduct extended synonym merging. In this step, the synthesis condition is initially parsed by LLM to extract all chemical substance names. Next, a pre-defined LLM prompt is used to ask GPT-4 to recognize all the extracted substance names and group them into identical substances. Finally, a reflection prompt is used to re-evaluate all the synonym pairs. The result is applied to merge the condition names together.

Third, using regular expressions, we follow-up to clean and format data containing special characters (spaces, commas, etc.). By doing this, we ensure the integrity and usability of synthesis

condition data.

Standardization of Numeric Conditions

We also standardize numerical data in time and temperature conditions. These data can have quality issues such as inconsistent unit and presence of special characters. The following treatments are then performed. First, using regular expressions, we extract and format relevant data such as time and temperature conditions. Second, we define the standard unit and convert each synthesis condition into it. For example, time conditions are converted into hours, temperatures are converted into Celsius. Room temperature is converted into 25°C uniformly.

Data Filtering by the Frequency of Synthesis Condition Values

In the application of MOFs property inference, the number of unique values on some condition such as organic linkers can be quite high due to their long-tailed distribution, which can lead to poor inference performance. In our experiment, we focus on the MOFs synthesized by top conditions. Particularly, we filter the data to only leave those synthesized from top-100 metal sources, top-100 organic linkers, and top-20 solvents, all after synonym merging. This filter setting can be adjusted for different application scenarios.

Feature Embedding for Metal, Organic Linker, and Solvent

The extracted synthesis conditions are embedded into vector representations for the follow-up machine learning, by augmenting their material/structural characteristics. The focused 10 synthesis conditions of a MOF are converted into a vector of length 206 by the following three steps.

First, using GPT-4, we obtain the chemical formulas and SMILES for the top-100 metal precursors and top-20 solvents of MOFs after synonym merging. For organic linkers, due to the complexity of their naming, GPT-4 can not obtain accurate SMILES. Therefore, we manually collect the SMILES for the top-100 organic linkers.

Second, based on the obtained SMILES, we use RDKit (43) to calculate the molecular features of metal precursors, organic linkers, and solvents, including their molecular weight, LogP values, the number of hydrogen bond donors and acceptors, Labute surface area, maximum molecular distance, molecular length, width, height, and topological polar surface area (TPSA).

Finally, using Pymatgen (44) and Matminer (45), we calculate several chemical features of metal salts, including oxidation states, elemental properties, atomic orbitals, electron affinity, and electronegativity differences. Additionally, we include features of the metal element contained in

the MOFs, such as its atomic mass, atomic radius, thermal conductivity, and detailed electronic configuration through vector representations.

Bayesian Optimization for Best Synthesis Route Discovery

To locate the best synthesis route of a MOF for optimal target property, we apply Bayesian optimization algorithm. The algorithm depends on an acquisition function to select the best synthesis route which are likely to yield significant improvements in the target value while also exploring regions with high uncertainty. Three acquisition functions are applied, each of which selects a separate best synthesis route. Let $f(x)$ represent the target property value (such as SA of a MOF) synthesized by route x . We seek to maximize $f(x)$, given a set of known synthesis routes x_i ($i = 1, \dots, N$) and their target property values $f(x_i)$. As the Bayesian algorithm requires to quantify the uncertainty for each inference, we apply the random forest (RF) model to predict $f(x)$ given N training data.

The first acquisition function, i.e., Expected Improvement (EI), quantifies the expected gain by evaluating the function at a new point in comparison to the current best observation.

$$EI(x) = \mathbb{E} [\max(0, f(x) - f(x^+))] \quad (6)$$

where $f(x^+)$ represents the observed maximal value of the function and $f(x)$ is the predicted value at x . In addition, Probability of Improvement (PI) and Upper Confidence Bound (UCB) acquisition functions are defined by

$$PI(x) = \Phi \left(\frac{\mu(x) - f(x^+)}{\sigma(x)} \right) \quad (7)$$

$$UCB(x) = \mu(x) + \kappa\sigma(x) \quad (8)$$

where μ and σ denote the expectation and standard deviation of the prediction at x , Φ is the CDF of standard normal distribution, and κ is a non-negative parameter balancing exploration and exploitation.

For all three acquisition functions, we create a grid of T candidate synthesis route cells in the input space ($T = 4M$ in this work and we have also tried sparser grid settings). The candidate with the largest acquisition function is selected as the best synthesis route. In future, we plan to also experiment with iterative Bayesian optimization. After each iteration, the suggested best route will be sent for lab experiment. The actually synthesized material and its measured target value will be used to update the inference model and execute a next iteration of Bayesian optimization.

References and Notes

1. A. C. Vaucher, *et al.*, Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications* **11** (1), 3601 (2020).
2. A. Trewartha, *et al.*, Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **3** (4) (2022).
3. J. Dagdelen, *et al.*, Structured information extraction from scientific text with large language models. *Nature Communications* **15** (1), 1418 (2024).
4. J. Choi, B. Lee, Accelerating materials language processing with large language models. *Communications Materials* **5** (1), 13 (2024).
5. T. He, *et al.*, Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chemistry of Materials* **32** (18), 7861–7873 (2020).
6. H. Park, Y. Kang, W. Choe, J. Kim, Mining insights on metal–organic framework synthesis from scientific literature texts. *Journal of Chemical Information and Modeling* **62** (5), 1190–1198 (2022).
7. L. T. Glasby, *et al.*, DigiMOF: a database of metal–organic framework synthesis information generated via text mining. *Chemistry of Materials* **35** (11), 4510–4524 (2023).
8. Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, O. M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society* **145** (32), 18048–18062 (2023).
9. K. N. Sasidhar, *et al.*, Enhancing corrosion-resistant alloy design through natural language processing and deep learning. *Science Advances* **9** (32), eadg7992 (2023).
10. W. Zhang, *et al.*, Fine-tuning large language models for chemical text mining. *Chemical Science* **15** (27), 10600–10611 (2024).
11. What is a MOF, MOF Commission of the International Zeolite Association, <https://www.iza-online.org/MOF/MOFforIZA.pdf>, retrieved on 2024-07-10.

12. L. Sun, *et al.*, Accelerated Dynamic Reconstruction in Metal–Organic Frameworks with Ligand Defects for Selective Electrooxidation of Amines to Azos Coupling with Hydrogen Production. *Angewandte Chemie* **136** (21), e202402176 (2024).
13. J. Wang, *et al.*, A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks. *Nature Communications* **15** (1), 1904 (2024).
14. H. B. Wu, X. W. Lou, Metal-organic frameworks and their derived materials for electrochemical energy storage and conversion: Promises and challenges. *Science Advances* **3** (12), eaap9252 (2017).
15. T. Xue, *et al.*, A customized MOF-polymer composite for rapid gold extraction from water matrices. *Science Advances* **9** (13), eadg4923 (2023).
16. A. H. Alawadhi, *et al.*, Harvesting Water from Air with High-Capacity, Stable Furan-Based Metal–Organic Frameworks. *Journal of the American Chemical Society* **146** (3), 2160–2166 (2024).
17. O. M. Yaghi, *et al.*, Reticular synthesis and the design of new materials. *Nature* **423** (6941), 705–714 (2003).
18. P. Chen, *et al.*, Machine-learning-guided morphology engineering of nanoscale metal-organic frameworks. *Matter* **2** (6), 1651–1666 (2020).
19. M. C. Swain, J. M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling* **56** (10), 1894–1904 (2016).
20. E. Kim, *et al.*, Machine-learned and codified synthesis parameters of oxide materials. *Scientific Data* **4** (1), 1–9 (2017).
21. T. Gupta, M. Zaki, N. A. Krishnan, Mausam, MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials* **8** (1), 102 (2022).

22. M. P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications* **15** (1), 1569 (2024).
23. S. Huang, J. M. Cole, A database of battery materials auto-generated using ChemDataExtractor. *Scientific Data* **7** (1), 260 (2020).
24. D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models. *Nature* **624** (7992), 570–578 (2023).
25. Z. Zheng, *et al.*, A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angewandte Chemie International Edition* **62** (46), e202311983 (2023).
26. Z. Zheng, *et al.*, Shaping the water-harvesting behavior of metal-organic frameworks aided by fine-tuned GPT models. *Journal of the American Chemical Society* **145** (51), 28284–28295 (2023).
27. Y. Kang, J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature Communications* **15** (1), 4705 (2024).
28. T. B. Brown, *et al.*, Language models are few-shot learners, in *Proceedings of the 34th International Conference on Neural Information Processing Systems* (2020), pp. 1877–1901.
29. P. Z. Moghadam, *et al.*, Development of a Cambridge Structural Database subset: a collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials* **29** (7), 2618–2625 (2017).
30. Clarivate, Web of Science platform, <https://clarivate.com/products/scientific-and-academic-research/> (2024).
31. GPT-4, OpenAI, <https://openai.com/index/gpt-4/>, retrieved on 2024-01-25.
32. H. Liu, *et al.*, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, in *Proceedings of the 36th International Conference on Neural Information Processing Systems* (2022), pp. 1950–1965.

33. M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938* (2023).
34. Y. Zhu, *et al.*, Large Language Models for Information Retrieval: A Survey. *arXiv preprint arXiv:2308.07107* (2023).
35. O. Ram, *et al.*, In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics* **11**, 1316–1331 (2023).
36. S. Robertson, H. Zaragoza, *et al.*, The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* **3** (4), 333–389 (2009).
37. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
38. N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
39. bert-base-uncased, Google, <https://huggingface.co/google-bert/bert-base-uncased/>, retrieved on 2023-10-08.
40. Y. Gao, *et al.*, Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
41. Q. Dong, *et al.*, A Survey for In-context Learning. *arXiv preprint arXiv:2301.00234* (2023).
42. J. White, *et al.*, A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).
43. G. Landrum, RDKit: Open-source cheminformatics, <https://www.rdkit.org>.
44. S. P. Ong, *et al.*, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013).
45. L. Ward, *et al.*, Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018).

46. H. Tian, W. Liu, other contributors, pdf2htmlEX, <https://github.com/pdf2htmlEX/pdf2htmlEX> (2024), accessed: 2024-07-18.
47. T. F. Willems, C. H. Rycroft, M. Kazi, J. C. Meza, M. Haranczyk, Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials* **149** (1), 134–141 (2012).
48. OpenAI, Prompt Engineering Guide, <https://platform.openai.com/docs/guides/prompt-engineering> (2024), retrieved on 2024-10-04.
49. S. M. Bsharat, A. Myrzakhan, Z. Shen, Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171* (2023).
50. P. Sahoo, *et al.*, A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927* (2024).

Acknowledgments

The authors would like to thank graduate students from School of Materials Science and Engineering, USTB, who helped to create the initial MOFs synthesis annotations on the CSD-MOFs dataset.

Funding: National Key R&D Program of China 2021YFB3500700 (LS, ZL, YY, WW, YZ, JL, SW, ZC, RL, NW, YL, ZL, HT, HG, GW)

NSFC Grant 62172026 (LS, YY, WW, ZC, NW, YL)

The Fundamental Research Funds for the Central Universities (LS, YY, WW, ZC, NW, YL)

State Key Laboratory of Complex & Critical Software Environment (LS, YY, WW, ZC, NW, YL)

Author contributions: Conceptualization: LS, GW

Machine learning and data analysis: ZL, YY, WW, HZ, ZC, NW, YL, YZ

System and visualization: YZ, SW, RL, ZL, HT

Material experiment: ZL, JL, HG, GW

Supervision: LS, HG, YZ, GW

Writing: LS and all co-authors

Competing interests: There are no competing interests to declare.

Data and materials availability: Key data of this study are provided in the supplemental data file. The uploaded data include the extracted synthesis conditions and the annotation data for all the three MOFs datasets, as well as input and output data for MOFs density and SA value inference. Other supporting data of this study have been deposited at https://github.com/BHT321/MOFs_Synthesis_Condition_Extraction/tree/main/Dataset. Our method and the obtained large-scale MOFs synthesis data is also available as an online executable engine (Fig. S11) and database (Fig. S10). (Supplementary Text) and the supplemental video provide more details. All codes for LLM synthesis extraction and microstructure property inference are available at: https://github.com/BHT321/MOFs_Synthesis_Condition_Extraction/tree/main/Code.

Supplementary materials

Supplementary Text

Figs. S1 to S19

Movie S1

Data S1

Supplementary Materials for LLM-based MOFs Synthesis Condition Extraction using Few-Shot Demonstrations

Lei Shi[†], Zhimeng Liu[†], Yi Yang, Weize Wu, Yuyang Zhang, Hongbo Zhang, Jing Lin, Siyu Wu,
Zihan Chen, Ruiming Li, Nan Wang, Yuankai Luo, Rui Wang, Zipeng Liu, Huobin Tan,
Hongyi Gao*, Yue Zhang*, Ge Wang*

*Corresponding author. Email: hygao@ustb.edu.cn, yue.zhang@wias.org.cn, gewang@ustb.edu.cn

[†]These authors contributed equally to this work.

This PDF file includes:

Supplementary Text

Figs. S1 to S19

Caption for Movie S1

Caption for Data S1

Other Supplementary Materials for this manuscript:

Movies S1

Data S1

Supplementary Text

MOFs Data

CSD, WoS, and the retrieved datasets

The primary data source of this work is the non-disordered MOF subset of Cambridge Structural Database (CSD) (29) retrieved in June 2022 (v5.43), which lists 84,898 MOFs covering the bonding motifs of all common MOFs in CSD (CSD-MOFs). The entry of a MOF in the database related to this work includes the MOF structure in CIF format, its physical properties, a DOI linking to the relevant publication, and a unique MOF ID.

The dataset is pre-processed according to the goal of this work. First, the full-text describing the MOFs under study should be available. Out of all the 84,898 MOFs, 78,741 have non-empty DOIs. Since the same DOI could be linked to multiple MOFs (one paper reporting more than one MOF), this leaves 39,579 different DOI links after deduplication and 36,177 downloadable paper full-texts. For the convenience of follow-up processing, we focus on the DOIs where the associated publication reports the information of only one MOF in CSD. This leads to a subset of 22,461 MOFs, each with a unique publication file in PDF format.

Next, the PDF of each MOF is converted to plain text (46) and segmented into paragraphs. The high-performance classification model in the main text is applied to detect synthesis paragraphs enclosing the desired synthesis condition information. Again, for the sake of convenience and accuracy, we only consider the 5,269 publications containing exactly one synthesis paragraph. Another 12,606 publications do not have any synthesis paragraph, probably because these papers are not related to MOFs experiments. The other 4,586 publications have more than one synthesis paragraph, as they are describing multiple MOFs or synthesis routes. Our pipeline could work with papers having more than one suite of synthesis conditions, but the potential MOF-synthesis mismatch may downgrade the application performance in evaluation. Therefore, throughout this work we stick to the core CSD-MOFs dataset with 5,269 publications/MOFs and their unique synthesis paragraphs.

To comprehensively evaluate our method and assess its application performance, we consider two additional data sources. The first focuses on the single type of UiO-66 MOF with the Zr metal. Synthesizing UiO-66 (Zr) MOFs using different conditions can result in varying structures and

material properties. Utilizing the Web of Science platform (30), we retrieved 4,524 unique research articles by searching the titles, abstracts, and keywords for “Zr, BDC” or “UiO, 66,” which forms the WoS-UiO-66 dataset. These papers encompass studies on both the modification and synthesis of UiO-66. We then employed a LLM to extract the surface area value of each UiO-66 MOF from the paper full-text. From 918 papers, we successfully extracted the surface area value. As we focus on MOFs synthesis and surface area inference of UiO-66, only papers reporting the synthesis of pure UiO-66 (the MOF without modification) are considered. Finally, the papers having exactly one synthesis paragraph, describing pure UiO-66, and presenting its surface area value, count to 261 articles, which is the core WoS-UiO-66 dataset. The last dataset is the SIMM data provided by Zhang et al. (10). The SIMM dataset contains 600 MOFs collected in Zheng et al. (8)’s work, whose synthesis routes are manually re-annotated in Zhang et al. (10) (as Zheng et al. (8) does not release the annotation data). Over the SIMM dataset, we keep 573 MOFs which has exactly one synthesis paragraph and one suite of synthesis route per MOF.

In addition, for material structure inference experiment, as the paper by Park et al. (6) does not release the original annotated training data, we are not able to obtain their extraction result on our full dataset for comparison. We then consider the extraction output provided by their paper on a subset of CSD-MOFs data (the KAIST dataset, containing 46701 MOFs). A data joint is performed on the KAIST dataset and our dataset with 5269 MOFs, resulting 814 MOFs as the evaluation data. The same data post-processings are conducted on both extraction results.

Microstructure Property Computation

For material evaluation purpose, we calculate structural and physical properties of the 5,269 MOFs in the core CSD-MOFs dataset. The CIF file of each MOF is retrieved from CSD and input to the Zeo++ tool (47). In total, four structural and physical properties are calculated: global cavity diameter, pore limiting diameter, largest cavity diameter, and framework density. We set the probe radius to 1.29\AA to simulate helium gas molecules, and the number of Monte Carlo samples to 100,000 to ensure the accuracy of calculations. All Zeo++ parameters adhere to standard routines, guaranteeing that the computed properties accurately represent the behavior of gas molecules within the MOF structure.

Data Annotations for Synthesis Paragraph and Condition

We invited eight experts on materials science and engineering to conduct the data annotations. Additionally, we developed a web-based interactive software to improve the efficiency of the annotation process (Fig. S2). On the synthesis paragraph annotation task, as the task is relatively easy, we use human annotation only. On the synthesis condition annotation task, which is more difficult, we apply a new method of human-AI interactive data curation, as detailed in the main text.

Synthesis Paragraph Annotation

440 papers were randomly selected from the CSD-MOFs dataset. Each paper was annotated by two different human experts, and in total 880 annotation tasks were assigned. Experts used our software shown in Fig. S2 to annotate synthesis-related paragraphs. After annotation, only paragraphs agreed by both labelers were considered valid, and the other disagreeing paragraphs were discarded.

This process yielded 1,349 valid synthesis paragraph annotations. To train the binary classification model, non-synthesis paragraphs are needed as negative samples. After removing all paragraphs annotated as synthesis paragraphs from each paper, the remaining paragraphs serve as negative samples. This method resulted in 11,783 negative samples used for training the synthesis paragraph detection model on the CSD-MOFs dataset. The WoS-UiO-66 dataset is processed by the same annotation procedure.

Synthesis Condition Annotation

On the CSD-MOFs dataset, we randomly selected 200 papers for synthesis condition annotation. The annotation process can be divided into five stages: task configuration, AI annotation, pilot annotation, batch annotation, and interactive data curation.

In the task configuration stage, domain experts define the key synthesis condition to be extracted and configure the annotation settings. The core objective is to maintain consistency for all ground-truth annotations. Due to the diversity of writing and representation styles in MOFs literature, consistent annotation helps to increase the frequency of major synthesis conditions, and therefore enhancing the effectiveness of follow-up machine learning based material inference. For example, we exclude MOF's activation conditions, which include activation temperature and activation time, after the pilot annotation process, as it is found that only few MOFs literature report activation

conditions. Also, molecular formulas for all synthesis conditions are excluded for annotation because the condition name serves as better predictor in material inference.

In the next stage, the GPT-4 LLM is applied on each synthesis paragraph for an initial synthesis condition annotation, i.e., the AI annotation stage (Fig. 3(I)~(II)). The annotation task configuration and material domain knowledge is feed in as the background prompt for LLM. As there are currently no ground-truth annotations, LLM works in a zero-shot mode. Although the performance of initial AI annotation is limited, this stage helps to resolve the fatigue issue of human labeler when working with a large amount of annotation tasks.

On the initial AI annotation, human labelers are instructed to apply best-effort annotations (Fig. 3(II)~(III)). Here two annotation stages are designed to maximize the accuracy of human annotation. First, *pilot annotation* on 20 papers of the CSD-MOFs dataset to validate and adjust the annotation task configuration. Each paper is annotated by two MOFs experts independently. The results on each paper are checked for their agreement between the two labelers. The expert labelers analyze and discuss any disagreement on the annotation, and resolve all the ambiguity and unclear issue during the annotation process. The pilot annotation stage also helps to optimize the design of our home-grown annotation software (Fig. S2). Second, *batch annotations* on the remaining 180 papers, handled by six human labelers in total. Again, each paper/paragraph is assigned two labelers for cross-check. After this annotation stage, the two labeling results on each paper are examined for intra-labeler agreement. For papers with higher degree of agreement than a pre-set threshold but not identical, another round of discussion is conducted to reach consensus on the final annotation. For papers with low degree of agreement, the paper/paragraph is simply discarded. The human annotation stages result in 147 papers with valid and consistent labels, and the other 53 papers are excluded for subsequent process.

Finally, we apply the human-AI interactive data curation (Fig. 3(III)~(IV)~(V)). First, the human annotation result is used as the demonstrations for a few-shot LLM extraction. Second, human labelers combine the few-shot LLM output and the last-stage human annotation result, known as the second-round human-AI data curation. In this way, both random errors induced by humans and common errors induced by AI due to the lack of specialized knowledge are largely resolved.

Synthesis Extraction and Structure Inference Performance on WoS-UiO-66 and SIMM Dataset

Few-shot LLM, as the core of our technical pipeline, exhibits excellent performance across different MOFs datasets. Apart from CSD-MOFs dataset reported in the main text, we also present additional results on another dataset, i.e., WoS-UiO-66 and SIMM dataset.

Prompt Engineering for Synthesis Extraction of UiO-66 MOFs

Similar to the synthesis extraction of CSD-MOFs dataset, we augment the domain knowledge of LLM through the background prompt. Due to differences between the datasets, prompts are modified to better meet the requirements. In the prompt, we explicitly require the LLM to focus on the synthesis of pure UiO-66 (also known as Zr-BDC) and ignore the other synthesis procedures like drying or crystallization process. Furthermore, we check whether our constraints are necessary for synthesis extraction of UiO-66. A preliminary extraction of UiO-66 synthesis conditions indicates that the LLM rarely makes errors in word sequence or modulator identification. Therefore, we remove some unnecessary constraints from our previous prompt, except for the clarification of synthesis and crystallization processes.

Synthesis Extraction Performance on WoS-UiO-66 Dataset

As mentioned in the main text, the few-shot examples used for our task are shown to be a critical factor to the synthesis extraction performance. Therefore, 87 UiO-66 MOFs with annotated paragraphs and synthesis conditions are used as examples. For synthesis extraction of UiO-66 MOFs, we employ the most efficient BM25 algorithm and a setting of $K = 6$ for the few-shot in-context learning.

In the comparison result of Fig. S3, it is shown that the few-shot LLM (Fig. S3(A)) achieves much higher macro-F1 (0.93) and ACC (0.90) than the zero-shot LLM (Fig. S3(B)), which achieves a macro-F1 score of 0.76 and ACC of 0.71. This result is similar to the extraction performance reported in the main text on the CSD-MOFs dataset. There are two detailed deficiency of the zero-shot LLM method. The first is that the zero-shot method can not distinguish between solvent and modulator, and it tends to classify some modulators (e.g., HAc or HCl) as solvents. The second is that the zero-shot method can not extract all the units of conditions. For example, if an amount is written as “30mg, 0.1mmol”, the LLM may extract either “30mg” or “0.1mmol”, but not both.

These two deficiencies of zero-shot method are addressed by the few-shot LLM by introducing domain-specific material knowledge. This demonstrates the universal advantage of our few-shot LLM method over the zero-shot LLM.

On the WoS-UiO-66 dataset, we also compare different RAG algorithms and the choice of number of shots (K). As shown in Fig. S4, the blue lines with circle symbols represent the performance variation with different K s in our task, using the best BM25 RAG algorithm. Both F1 and ACC improve most significantly from zero-shot to one-shot, with an increase of more than 0.1. This improvement is greater than the results on the CSD-MOFs dataset, possibly because of the lack of specialized knowledge on UiO-66 in the LLM. Provided with high-quality UiO-66 material knowledge, LLM is able to achieve the same extraction performance obtained on the CSD-MOFs dataset. The performance continues to improve until $K = 6$. After this peak (F1=0.93, ACC=0.90), the metrics fluctuate without surpassing the best performance. Meanwhile, the few-shot method with random example selection (orange lines), BERT algorithm (green lines) and SBERT (red lines) exhibit the same trend as K increases. The performance of both BERT and random example selection consistently fall behind that of the BM25 algorithm, while SBERT and BM25 perform almost identically. The variation among these four algorithms is smaller than the results on the CSD-MOFs dataset, possibly due to the higher similarity among different UiO-66 synthesis routes. Because a lot of synthesis conditions used for UiO-66 synthesis are usually the same, especially the precursor names, the overall synthesis context and paragraphs are highly similar to each other. Therefore, retrieving different synthesis paragraphs as examples does not significantly impact the synthesis extraction performance on the WoS-UiO-66 MOFs dataset.

Structure Inference Performance on UiO-66

We also design a material inference task to compare the performance of few-shot vs. zero-shot synthesis extraction methods. A key structure property of MOFs, i.e., the surface area, is inferred in this task, using the extracted synthesis conditions including metals, organic linkers, solvents, modulators, and reaction duration/temperature. The input data is the extraction result on the WoS-UiO-66 dataset. They are post-processed using the same approach applied to the CSD-MOFs dataset. After filtering out sparsely appearing synthesis conditions, we have a smaller dataset in both few-shot and zero-shot extraction result, with 151 and 85 MOFs respectively. A 19-dimensional embedding vector is obtained for each MOF's synthesis conditions.

The random forest model is selected for the inference as it supports quantifying the uncertainty of prediction. We compare few-shot LLM and zero-shot LLM in a 10-fold cross-validation. The inference performance is reported by the coefficient of determination (R^2). It is shown that the synthesis conditions obtained by the few-shot method exhibit significantly higher predictive power than those obtained by the zero-shot method. The R^2 value of few-shot is 0.1404 while the zero-shot is 0.0751. This inference performance is lower than the prediction of MOF's microstructural property on the CSD-MOFs dataset, mainly because we have less data on UiO-66. The inference model is further utilized in suggesting the best synthesis conditions for optimal surface area value.

Synthesis Extraction Performance on SIMM Dataset

In addition to the WoS-UiO-66 dataset, we also compare the performance of Few-shot RAG algorithm with zero-shot LLM extraction on SIMM dataset. We apply the Human-AI annotation process on the origin SIMM training and testing datasets, resulting in 573 MOFs for further extraction. For the synthesis extraction of SIMM MOFs, we employ the BM25 algorithm with a setting of $K = 4$ for few-shot in-context learning.

As shown in the comparison results of Fig. S5, the few-shot LLM achieves significantly higher macro-F1 (0.96) and ACC (0.94) compared to the zero-shot LLM, which achieves a macro-F1 score of 0.74 and ACC of 0.70.

MOFs Synthesis Extraction Engine, Database, and Visualization

We developed an online system using the method proposed in this paper, which is capable of automatically executing the LLM-based synthesis extraction and supporting the visual analysis and database retrieval of the extraction result on the CSD-MOFs dataset.

Online MOFs Synthesis Extraction Engine

To streamline the entire synthesis extraction workflow, allow easy access to our method, and support online usage of the engine over any related literature, we developed the MOFs Synthesis Condition Extraction Engine.

The engine is developed using a frontend-backend separation architecture. The frontend follows the MVC design pattern and is built using the well-established open-source Vue 3 framework, TypeScript scripting language, and the Vuetify component library. Data is retrieved by sending Axios requests to the backend. The backend adopts a layered architecture and leverages the Spring

Boot 3 framework to handle requests. High-performance paragraph synthesis extraction models and synthesis condition extraction models are exposed via FastAPI endpoints, and the backend communicates with these services using FeignClient to send requests and process the returned results.

The system utilizes MySQL as the relational database and MinIO for file storage services. All data related to the synthesis extraction workflow is stored in the databases. The data manipulation and connection operations are managed via Spring.

MOFs Synthesis Database

Using the synthesis paragraph detection algorithm, we processed 36,177 papers from the CSD-MOFs dataset and extracted 57,081 synthesis paragraphs, on which we then executed the proposed few-shot LLM extraction method. The large amount of extraction results are stored in the MySQL database. Figure S6 gives the basic statistics of the database. The database is mainly composed of tables on the 10 extracted synthesis conditions and the original synthesis paragraph, mostly with over 10k records. We implemented the faceted search capability on the database. As shown in Fig. S10, the visual user interface allows to search on any retrieved synthesis conditions as well as the original literature metadata and the synthesis paragraph content. Fig. S7 gives the page of search result upon the basic search. By the Elasticsearch technology, the system also supports advanced search that combines multiple queries via boolean logic operators (Fig. S8). Typical advanced search result visualization is given in Fig. S9.

User Interface for Online MOFs Synthesis Extraction and LLM Output Visualization

We integrate the entire workflow of MOFs synthesis extraction from scientific literature into the same interactive user interface. The user operation on the interface is composed of the following steps:

First, users upload one or multiple papers related to the MOFs synthesis to our system (Fig. S11). The paper files will be automatically converted into textual file format suitable for the follow-up synthesis condition extraction.

Second, after uploading and converting paper files, the system will automatically detect synthesis paragraphs from all the uploaded papers. A literature status list is displayed to show the state of synthesis paragraph detection on all uploaded papers (Figure S12).

Third, users can view the detected synthesis paragraphs and configure the algorithm parameters

for the follow-up synthesis condition extraction (Figure S13). At the top of the interface, a drop-down list allows to select a specific paper from the current batch. The left panel will displays a preview of the selected literature. The central panel shows the detected synthesis paragraphs, allowing users to choose a few paragraphs to process. In the right part, a configuration panel allows users to set the current algorithm parameters for the synthesis extraction, which include the choice of LLM model and RAG algorithm, as well as the number of few-shots (K).

LLM Output Visualization

The system supports visualization and interactive analysis of the synthesis extraction result given by LLM. In the visualization interface shown in Fig. S14, (A) gives the four overall performance metrics for the synthesis extraction. (B) expands the performance metrics on each of the 10 synthesis conditions, using a heat map display. (C) provides the detailed view of synthesis condition extraction, showing both the extracted entities and the ground-truth annotations for comparison.

The interface can be switched to the second tab view (Figure S15) to drill down to more detailed performance on each uploaded literature. In a third tab view (Figure S16), the 2D projection visualization shows the distribution of extracted 10 synthesis conditions. Here, red dots indicate newly extracted suites of synthesis conditions. Blue dots represent the corresponding ground-truth synthesis conditions by annotation. Black dots indicates the other synthesis conditions already in the database, showing a typical synthesis condition distribution.

Database Retrieval

The system also supports the visualization of database retrieval. As shown in Figure S17, users can view the distribution of retrieved synthesis conditions through a 2D scatterplots. The detailed information of each synthesis condition is displayed in the list below.

LLM Prompts For Synthesis Extraction

Introduction of LLM Prompt Structure

This section presents the structure and specific content of the prompts used for synthesis condition extraction. In our implementation, we use the OpenAI API as the GPT calling tool, which differs slightly in its display and usage compared to the web version. Without otherwise noted, we use the GPT-4 turbo model.

An example of a one-shot prompt for extracting MOFs is shown below:

Prompt Demonstration

Prompt:

I require your assistance in efficiently and precisely extracting synthesis parameters from chemical literature...

Sample Input 1:

2.2.1. Synthesis of $[\text{Cu I} (4,4' \text{-bpy})]_2 [\text{H}_2\text{SiW}_12\text{O}_{40}] \cdot 3\text{H}_2\text{O}$ (1). $\text{H}_4\text{SiW}_12\text{O}_{40} \cdot 18\text{H}_2\text{O}$ [28] (1.15 g, 0.356 mmol), $\text{Cu}(\text{Ac})_2 \cdot 2\text{H}_2\text{O}$ (0.16 g, 0.7 mmol), 4,4' -bipy (0.05 g, 0.32 mmol), and en (0.25 g, 4.16 mmol) were dissolved in 10 mL water...

Sample Output 1:

```
[  
 {  
   "Compound_Name": [  
     "[Cu I (4,4' -bpy)]_2 [H_2 SiW_12 O_40] · 3H_2 O"  
   ],  
   "Metal_Source": [  
     {  
       "precursor_name": "H_4 SiW_12 O_40 · 18H_2 O",  
       "amount": "1.15 g, 0.356 mmol"  
     },  
     {  
       "precursor_name": "Cu(Ac)_2 · 2H_2 O",  
       "amount": "0.16 g, 0.7 mmol"  
     }  
   ],  
   ...  
 }
```

Input:

Synthesis of [Cd(L)(BIB) 0.5] n (1) A mixture of H 2 L (0.2703 g, 1 mmol), Cd(OAc) 2 · 2H 2 O (0.2673 g, 1 mmol), NaOH (0.0801 g, 2 mmol) and BIB (0.189 g, 1 mmol) in 10 mL of distilled water was sealed in a 23 mL vial, and then heated to 150 °C for 5 days.

Output:

```
[  
 {  
   "Compound_Name": [  
     "[Cd(L)(BIB) 0.5 ] n"  
   ],  
   "Metal_Source": [  
     {  
       "precursor_name": "Cd(OAc) 2 · 2H 2 O",  
       "amount": "0.2673 g, 1 mmol"  
     }  
   ],  
   ...  
 }
```

The above box illustrates the workflow of our GPT-based extraction process, divided into five sections: “Prompt,” “Sample Input,” “Sample Output,” “Input,” and “Output.” The “Prompt” describe the task for the model and provide necessary information. The “Sample Input” and “Sample Output” provide examples of input texts and the desired output format, which help guide the model’s responses. The “Input” section contains the actual synthesis paragraph to be processed, and the “Output” section presents the GPT-generated extraction results.

The prompt structure when programmatically invoking GPT differs from that of the web version, and an example implementation is shown in the box below, which displays the actual code used for invocation and the method for obtaining responses. The complete code can be found on GitHub: https://github.com/passingby000/MOFs_Synthesis_Condition_Extraction

GPT Calling Code in Python

```
response = client.chat.completions.create(  
    model="gpt-4-turbo",  
    messages=[  
        {"role": "system", "content": Prompt},  
        {"role": "user", "content": Sample_Input_1},  
        {"role": "assistant", "content": Sample_Output_1},  
        {"role": "user", "content": Input}  
    ],  
    temperature=0  
)  
  
Output = response.choices[0].message.content.strip()  
return Output
```

For clearer and more intuitive demonstration, the prompt figures use “Prompt” to represent the “system” in code, “Sample Input 1” to indicate the first “user”, “Sample Output 1” to indicate the first “assistant”. Each “user-assistant” pair acts as a sample input/output pair. The prompt figure use “Input” to represent the final “user” containing the synthesis paragraph for extraction, from which the model generates an output. Models are set according to the task, and temperature is set to 0 to keep the stability of output. The GPT-API “Output” is retrieved and parsed from the API model response.

Our prompt structure, depicted in Fig. S18, consists of five functional components, including role establishment, task definition, background, mission details, and structure sample. For brevity, some specific content has been omitted. The rationale behind our prompt design is detailed in the subsequent sections of the main text. The design references several related works (8), including the official OpenAI guidelines (48), and incorporates the latest paradigms in prompt engineering (42, 49, 50). As a comparative study, the prompt structure used by Zheng et al. is shown in Fig. S19. We have similarly adopted their method for eliminating hallucinations and added background knowledge and constraints specific to our task to enhance the model’s performance on the CSD-

MOFs and WoS-UiO-66 datasets.

In summary, the carefully designed prompt structure and API interaction are crucial for accurately and efficiently extracting synthesis conditions of MOFs from chemical literature using GPT models.

Prompts for Zero-Shot LLM in CSD-MOF Experiment

Zero-shot LLM extraction was conducted over the 5,269 filtered synthesis paragraphs from the CSD-MOFs dataset, using the GPT-4 API. The extracted synthesis conditions were then cleaned and utilized in subsequent experiments.

The following table presents the prompts, processes, and sample results used in this extraction task. The DOI of following condition extraction example paper is [10.1134/S1063774518040296](https://doi.org/10.1134/S1063774518040296)

Zero-shot Prompts for CSD-MOF

Prompt:

You are a chemical expert with 20 years of experience in reviewing literature and extracting key information. Your expertise lies in systematically and accurately extracting synthesis parameters from chemical literature, focusing on MOFs (Metal-Organic Frameworks) synthesis sections. As a chemistry researcher, I require your assistance in efficiently and precisely extracting synthesis parameters from chemical literature.

Your task is to summarize the following details for a JSON format table from some input: ‘Compound_Name’, ‘Metal_Source’, ‘Organic_Linkers’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, ‘Reaction_Temperature’. Among them, ‘Metal_Source’, ‘Organic_Linkers’, ‘Solvent’, and ‘Modulator’ should also contain their amounts.

Only focus on the MOF synthesis process. Ignore other processes, including the organic precursor synthesis pre-process and the active and crystallization post-process, if mentioned in the paragraph.

The detailed format descriptions for each class are below:

The output should be a JSON table list. Each JSON format table represents a MOF.

If there is only one MOF, the JSON list should only have one JSON format table. If there is more than one MOF, they should be put in different JSON tables.

Each JSON format table should contain: “Compound_Name”, “Metal_Source”, “Organic_Linked”, “Solvent”, “Modulator”, “Reaction_Time”, “Reaction_Temperature”.

The “Compound_Name”, “Reaction_Time”, “Reaction_Temperature” should be lists of strings.

The “Metal_Source”, “Organic_Linked”, “Solvent”, and “Modulator” should all be lists of dicts. The dict should have the keys “precursor_name” and “amount”. “amount” includes weight, volume, or molar weight, presented according to the paragraph text, but not including the concentration, proportion, or rate.

The “Compound_Name” should contain the full MOF chemical name that has been synthesized according to the paragraph text. Extract the suffix like “}n” or “[n”, because it represents the structure of the compound, acting as part of the chemical name. Remove the prefix which is not part of the chemical MOF name like “compound” or “complex” and the suffix like “(1)” or “(2)”.

Extract the name and amount of “Metal_Source”, “Organic_Linked”, “Solvent”, and “Modulator” during the synthesis process of MOF according to the paragraph, and put them in the dict with the keys “precursor_name” and “amount”.

The “Reaction_Time”, “Reaction_Temperature” should be extracted from the paragraph. If there is no time or temperature in the text, set the value to be an empty list.

If some values are missing, not presented, or not mentioned, keep them empty in the output JSON text. The output should only contain the list.

The output structure should be like this:

```
[  
{  
    "Compound_Name": [] ,  
    "Metal_Source": [
```

```
{  
    "precursor_name": "",  
    "amount": ""  
},  
],  
"Organic_Linker": [  
    {  
        "precursor_name": "",  
        "amount": ""  
    },  
    {  
        "precursor_name": "",  
        "amount": ""  
    }  
],  
"Solvent": [  
    {  
        "precursor_name": "",  
        "amount": ""  
    },  
    {  
        "precursor_name": "",  
        "amount": ""  
    }  
],  
"Modulator": [  
    {  
        "precursor_name": "",  
        "amount": ""  
    }  
]
```

```
        },
        ],
        "Reaction_Time": [],
        "Reaction_Temperature": []
    }
]
```

Input:

Synthesis of $[Cd(L)(BIB)0.5]n$ (1) A mixture of H_2L (0.2703 g, 1 mmol), $Cd(OAc)_2 \cdot 2H_2O$ (0.2673 g, 1 mmol), NaOH (0.0801 g, 2 mmol) and BIB (0.189 g, 1 mmol) in 10 mL of distilled water was sealed in a 23 mL vial, and then heated to 150°C for 5 days. Colorless block crystals of 1 were collected in 56% yield based on Cd. Anal. calcd. (%) for $C_{20}H_{15}CdN_2O_5$: C, 50.49; H, 3.18; N, 5.89. Found (%): C, 50.45; H, 3.15; N, 5.82.

Output:

```
[
{
    "Compound_Name": [
        "[Cd(L)(BIB)0.5]n"
    ],
    "Metal_Source": [
        {
            "precursor_name": "Cd(OAc)2 · 2H2O",
            "amount": "0.2673 g"
        }
    ],
    "Organic_Linker": [
        {
            "precursor_name": "H2L",
            "amount": "0.2703 g"
        }
    ]
}]
```

```

    },
    {
        "precursor_name": "BIB",
        "amount": "0.189 g"
    }
],
"Solvent": [
    {
        "precursor_name": "distilled water",
        "amount": "10 mL"
    }
],
"Modulator": [
    {
        "precursor_name": "NaOH",
        "amount": "0.0801 g"
    }
],
"Reaction_Time": [
    "5 days"
],
"Reaction_Temperature": [
    "150°C"
]
}
]

```

Prompts for Few-Shot LLM in CSD-MOF Experiment

Few-shot LLM extraction was conducted over the 5,269 filtered synthesis paragraphs from the CSD-MOFs dataset, using the GPT-4 API. To select 3 relevant examples from 137 annotations, we

employed the BM25 algorithm. These examples were then used to extract synthesis conditions. To specifically improve extraction quality, background information and constraints based on domain knowledge were added to the prompts. The results were subsequently cleaned and utilized in subsequent experiments.

The following table presents the prompts, examples, processes, and sample results used in this extraction task. The DOI of following condition extraction example paper is [10.1134/S1063774518040296](https://doi.org/10.1134/S1063774518040296).

Few-shot Prompts for CSD-MOF

Prompt:

You are a chemical expert with 20 years of experience in reviewing literature and extracting key information. Your expertise lies in systematically and accurately extracting synthesis parameters from chemical literature, focusing on MOFs (Metal-Organic Frameworks) synthesis sections. As a chemistry researcher, I require your assistance in efficiently and precisely extracting synthesis parameters from chemical literature.

Your task is to summarize the following details for a JSON format table from some input: ‘Compound_Name’, ‘Metal_Source’, ‘Organic_Linkers’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, ‘Reaction_Temperature’. Among them, ‘Metal_Source’, ‘Organic_Linkers’, ‘Solvent’, and ‘Modulator’ should also contain their amounts.

Only focus on the MOF synthesis process. Ignore other processes, including the organic precursor synthesis pre-process and the active and crystallization post-process, if mentioned in the paragraph.

Background Information and Detailed Instructions:

Compound_Name of MOFs (Metal-Organic Frameworks): MOFs are porous materials formed by the coordination of metal ions or clusters with organic ligands. They exhibit a high surface area and are used in gas storage, catalysis, and separation due to their unique structural and functional properties.

Metal_Source: In MOF synthesis, a Metal Source is a precursor compound containing

metal ions that form part of the MOF structure. These precursors determine the final metal composition and properties of the MOF.

Organic Linker: Refers to the organic precursor molecule linking metal ions or clusters in the MOF, influencing the framework's topology, porosity, and functionality.

Solvent: The liquid medium in which reactants are dissolved.

Modulator: The modulator aims to adjust the reaction condition, such as PH value.

Reaction process: The synthesis process that produces the MOF materials.

The detailed format descriptions for each class are below:

The output should be a JSON table list. Each JSON format table represents a MOF.

If there is only one MOF, the JSON list should only have one JSON format table. If there is more than one MOF, they should be put in different JSON tables.

Each JSON format table should contain: "Compound_Name", "Metal_Source", "Organic_Linker", "Solvent", "Modulator", "Reaction_Time", "Reaction_Temperature".

The "Compound_Name", "Reaction_Time", "Reaction_Temperature" should be lists of strings.

The "Metal_Source", "Organic_Linker", "Solvent", and "Modulator" should all be lists of dicts. The dict should have the keys "precursor_name" and "amount". "amount" includes weight, volume, or molar weight, presented according to the paragraph text, but not including the concentration, proportion, or rate.

The "Compound_Name" should contain the full MOF chemical name that has been synthesized according to the paragraph text. Extract the suffix like "}n" or "]n", because it represents the structure of the compound, acting as part of the chemical name. Remove the prefix which is not part of the chemical MOF name like "compound" or "complex" and the suffix like "(1)" or "(2)".

Extract the name and amount of "Metal_Source", "Organic_Linker", "Solvent", and "Modulator" during the synthesis process of MOF according to the paragraph, and put them in the dict with the keys "precursor_name" and "amount".

The "Reaction_Time", "Reaction_Temperature" should be extracted from the paragraph. If

there is no time or temperature in the text, set the value to be an empty list.

If some values are missing, not presented, or not mentioned, keep them empty in the output JSON text. The output should only contain the list.

The output structure should be like this:

```
[  
  {  
    "Compound_Name": [],  
    "Metal_Source": [  
      {  
        "precursor_name": "",  
        "amount": ""  
      }  
    ],  
    "Organic_Linkers": [  
      {  
        "precursor_name": "",  
        "amount": ""  
      },  
      {  
        "precursor_name": "",  
        "amount": ""  
      }  
    ],  
    "Solvent": [  
      {  
        "precursor_name": "",  
        "amount": ""  
      },  
      {  
        "precursor_name": "",  
        "amount": ""  
      }  
    ]  
  }]
```

```

{
    "precursor_name": "",
    "amount": ""

},
"Modulator": [
{
    "precursor_name": "",
    "amount": ""

},
],
"Reaction_Time": [],
"Reaction_Temperature": []
}
]

```

Sample Input 1:

2.2.1. Synthesis of $[\text{Cu I} (4,4' \text{-bpy})]_2 [\text{H}_2\text{SiW}_12\text{O}_{40}] \cdot 3\text{H}_2\text{O}$ (1). $\text{H}_4\text{SiW}_12\text{O}_{40} \cdot 18\text{H}_2\text{O}$ [28] (1.15 g, 0.356 mmol), $\text{Cu}(\text{Ac})_2 \cdot 2\text{H}_2\text{O}$ (0.16 g, 0.7 mmol), 4,4'-bipy (0.05 g, 0.32 mmol), and en (0.25 g, 4.16 mmol) were dissolved in 10 mL water. The pH of the resulting mixture was adjusted to 2.7 by adding HCl solution, and then the mixture was stirred for 30 min in air. The final solution was transferred into a 25 mL Teflon-lined autoclave at 170 °C for 3 days. Then the autoclave was cooled at 10 °C/h to room temperature. The resulting dark red block crystals were filtered off, washed with distilled water, and dried in a desiccator at room temperature (38% yield based on W).

Sample Output 1:

```
[
{
    "Compound_Name": [

```

```
"[Cu I (4,4' -bpy)] 2 [H 2 SiW 12 O 40 ] · 3H 2 O"  
],  
"Metal_Source": [  
    {  
        "precursor_name": "H 4 SiW 12 O 40 · 18H 2 O",  
        "amount": "1.15 g, 0.356 mmol"  
    },  
    {  
        "precursor_name": "Cu(Ac) 2 · 2H 2 O",  
        "amount": "0.16 g, 0.7 mmol"  
    }  
],  
"Organic_Linker": [  
    {  
        "precursor_name": "4,4' -bipy",  
        "amount": "0.05 g, 0.32 mmol"  
    },  
    {  
        "precursor_name": "en",  
        "amount": "0.25 g, 4.16 mmol"  
    }  
],  
"Solvent": [  
    {  
        "precursor_name": "water",  
        "amount": "10 mL"  
    }  
],  
"Modulator": [
```

```

    {
        "precursor_name": "HCl solution",
        "amount": ""

    },
    "Reaction_Time": [
        "3 days"
    ],
    "Reaction_Temperature": [
        "170 °C"
    ]
}
]

```

Sample Input 2:

2.2. Preparation of μ -o-[Ni 2 (4,4' -bpy) 2 (oba) 2] n (1) NiSO₄ · 6H₂O (0.1314 g, 0.5 mmol), H₂ oba (0.1291 g, 0.5 mmol) and 4,4' -bipyridine (0.0781 g, 0.5 mmol) were dissolved in 10 mL distilled water, 150 mL triethylamine was added slowly to the mixture upon stirring, after further 20 min continuous stirring the final mixture was sealed in a 23 mL Teflon-lined stainless vessel, and heated at 160 °C for 72 h.

Sample Output 2:

```

[
{
    "Compound_Name": [
        "\text{\mu}-o-[Ni 2 (4,4' -bpy) 2 (oba) 2 ] n"
    ],
    "Metal_Source": [
        {
            "precursor_name": "NiSO 4 · 6H 2 O",

```

```
        "amount": "0.1314 g, 0.5 mmol"
    }
],
"Organic_Linker": [
{
    "precursor_name": "H 2 oba",
    "amount": "0.1291 g, 0.5 mmol"
},
{
    "precursor_name": "4,4' -bipyridine",
    "amount": "0.0781 g, 0.5 mmol"
}
],
"Solvent": [
{
    "precursor_name": "water",
    "amount": "10 mL"
}
],
"Modulator": [
{
    "precursor_name": "triethylamine",
    "amount": "150 mL"
}
],
"Reaction_Time": [
    "72 h"
],
"Reaction_Temperature": [
```

```
    "160 °C"
]
}
]
```

Sample Input 3:

Synthesis of $[\text{Ag}(\text{H}_3\text{bptc})(\text{bpe})]\cdot 2\text{H}_2\text{O}$ n (1) Compound 1 was obtained by reaction of AgNO_3 (0.2 mmol), bpe (0.05 mmol) and H_4bptc (0.05 mmol) in 4:1:1 molar ratio in 15 mL of water and 1 mL of triethylamine under hydrothermal conditions (at 150 °C for 6 days and cooled to room temperature with a 5 °Ch 1 rate).

Sample Output 3:

```
[
{
  "Compound_Name": [
    "{[\text{Ag}(\text{H}_3\text{bptc})(\text{bpe})]\cdot 2\text{H}_2\text{O}}\text{ n}"
  ],
  "Metal_Source": [
    {
      "precursor_name": "AgNO 3",
      "amount": "0.2 mmol"
    }
  ],
  "Organic_Linker": [
    {
      "precursor_name": "bpe",
      "amount": "0.05 mmol"
    },
    {
      "precursor_name": "H_4 bptc",
      "amount": "0.05 mmol"
    }
  ]
}]
```

```
        "amount": "0.05 mmol"
    }
],
"Solvent": [
{
    "precursor_name": "water",
    "amount": "15 mL"
},
{
    "precursor_name": "triethylamine",
    "amount": "1 mL"
}
],
"Reaction_Time": [
    "6 days"
],
"Reaction_Temperature": [
    "150 °C"
]
}
```

Input:

Synthesis of [Cd(L)(BIB) 0.5] n (1) A mixture of H₂L (0.2703 g, 1 mmol), Cd(OAc)₂ · 2H₂O (0.2673 g, 1 mmol), NaOH (0.0801 g, 2 mmol) and BIB (0.189 g, 1 mmol) in 10 mL of distilled water was sealed in a 23 mL vial, and then heated to 150 °C for 5 days. Colorless block crystals of 1 were collected in 56% yield based on Cd.

Output:

```
[
```

```
{  
    "Compound_Name": [  
        "[Cd(L)(BIB) 0.5 ] n"  
    ],  
    "Metal_Source": [  
        {  
            "precursor_name": "Cd(OAc) 2 + 2H 2 O",  
            "amount": "0.2673 g, 1 mmol"  
        }  
    ],  
    "Organic_Linker": [  
        {  
            "precursor_name": "H 2 L",  
            "amount": "0.2703 g, 1 mmol"  
        },  
        {  
            "precursor_name": "BIB",  
            "amount": "0.189 g, 1 mmol"  
        }  
    ],  
    "Solvent": [  
        {  
            "precursor_name": "water",  
            "amount": "10 mL"  
        }  
    ],  
    "Modulator": [  
        {  
            "precursor_name": "NaOH",  
            "amount": "0.5 g, 0.01 mol"  
        }  
    ]  
}
```

```
        "amount": "0.0801 g, 2 mmol"  
    },  
    ],  
    "Reaction_Time": [  
        "5 days"  
    ],  
    "Reaction_Temperature": [  
        "150 °C"  
    ]  
}  
]
```

Prompts for Zero-Shot LLM in WoS-UiO-66 Experiment

Zero-shot LLM extraction was conducted over the 261 filtered synthesis paragraphs from the WoS-UiO-66 dataset, using the GPT-4 API. These paragraphs were then cleaned and utilized in subsequent experiments.

Since this experiment was conducted on a pure UiO-66 dataset, relevant background and requirements were added to the prompts to determine whether the material is indeed pure UiO-66. The model was also instructed to exclude the synthesis of modified processes, or unrelated materials.

The following table presents the prompts, processes, and sample results used in this extraction task. The DOI of following condition extraction example paper is [10.1016/j.ces.2019.04.006](https://doi.org/10.1016/j.ces.2019.04.006)

Zero-shot Prompts for WoS-UiO-66

Prompt:

You are a chemical expert with 20 years of experience in reviewing literature and extracting key information. Your expertise lies in systematically and accurately extracting synthesis parameters from chemical literature, specifically focusing on the synthesis sections of Metal-Organic Frameworks (MOFs). As a chemistry researcher, I require your assistance

in efficiently and precisely extracting the synthesis parameters of UIO-66, a type of MOF, from chemical literature. I will tip 10\$ for more precise extraction result.

Your task is to summarize the following details into a JSON format table from the input paragraph: ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, and ‘Reaction_Temperature’. ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, and ‘Modulator’ should also include their respective amounts. Only focus on the synthesis process. Ignore other processes, including the organic precursor synthesis pre-process, the drying process, the crystallization post-process, and any modification processes if mentioned. Do not extract them.

Only extract details for pure UIO-66, also known as pure Zr-BDC. Do not extract information on other chemical compounds such as UIO-67. Pure UIO-66 refers to the unmodified compound; therefore, do not extract information on UIO-66-NH₂ or Cu@UIO-66, as they are not considered pure.

Detailed Format Descriptions and Requirements for Each Class:

The output should be a JSON table list. Each JSON format table represents a UIO-66 synthesis process. If there is only one synthesis process, the JSON list should contain only one JSON format table. If there is more than one synthesis process, they should be placed in separate JSON tables.

Each JSON format table should contain: ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, and ‘Reaction_Temperature’.

- ‘Reaction_Time’ and ‘Reaction_Temperature’ should be lists of strings.
- ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, and ‘Modulator’ should all be lists of dictionaries. Each dictionary should have the keys ‘precursor_name’ and ‘amount’. ‘amount’ includes weight, volume, or molar weight, presented according to the paragraph text, but not including the concentration, proportion, or rate. Do not include any proportions, ratios, concentrations, or percentages in the name or amount of any parameter.

Extract the name and amount of ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, and ‘Modulator’ during the synthesis process of MOF according to the paragraph, and place them in the dictionary with the keys ‘precursor_name’ and ‘amount’.

‘Reaction_Time’ and ‘Reaction_Temperature’ should be extracted from the paragraph. If there is no time or temperature mentioned in the text, set the value to an empty list.

If some values are missing, not presented, or not mentioned, leave them empty in the output JSON text. The output should only contain the list.

Example Output Structure:

```
[  
  {  
    "Metal_Source": [  
      {  
        "precursor_name": "",  
        "amount": ""  
      }  
    ],  
    "Organic_Linker": [  
      {  
        "precursor_name": "",  
        "amount": ""  
      },  
      {  
        "precursor_name": "",  
        "amount": ""  
      }  
    ],  
    "Solvent": [
```

```
{  
    "precursor_name": "",  
    "amount": ""  
},  
{  
    "precursor_name": "",  
    "amount": ""  
}  
,  
]  
,  
"Modulator": [  
    {  
        "precursor_name": "",  
        "amount": ""  
    }  
,  
]  
,  
"Reaction_Time": [],  
"Reaction_Temperature": []  
}  
]
```

Input:

2.1.2. Synthesis UiO-66 was synthesized according to a slightly modified protocol based on previous report by Schaate (2011). In detail, ZrCl₄ (1.00 g, 4.29 mmol) and acetic acid (7.36 mL, 128.7 mmol) were dissolved in DMF (250 mL, 3.23 mol) by ultrasound for about 10 min, which was then supplemented with H₂BDC (0.713 g, 4.29 mmol). The solution was transferred to a 500 mL Teflon-lined autoclave and maintained at 393 K under static conditions for 24 h. The precipitates were isolated and then subject to three-time solvent exchange using fresh 30 mL DMF, in which the precipitates were soaked in DMF for 12h

each time. Afterwards, the precipitates were soaked in free ethanol (30 mL) for three-time solvent exchange similarly. Finally, the solids were first dried in an oven at 80°C for 12h, and then dried under vacuum at 453K for 12h. UiO-67 was synthesized using a previously reported method in literature (Oien-Odegaard et al., 2016) with a slight modification.

Output:

```
[  
  {  
    "Metal_Source": [  
      {  
        "precursor_name": "ZrCl4",  
        "amount": "1.00 g, 4.29 mmol"  
      },  
      ],  
    "Organic_Linker": [  
      {  
        "precursor_name": "H2BDC",  
        "amount": "0.713 g, 4.29 mmol"  
      },  
      ],  
    "Solvent": [  
      {  
        "precursor_name": "DMF",  
        "amount": "250 mL, 3.23 mol"  
      },  
      {  
        "precursor_name": "acetic acid",  
        "amount": "7.36 mL, 128.7 mmol"  
      }  
    ]  
  }]
```

```

],
"Modulator": [
{
  "precursor_name": "Acetic acid",
  "amount": "7.36 mL"
},
],
"Reaction_Time": [
  "24 hours"
],
"Reaction_Temperature": [
  "393 K"
]
}
]

```

Prompts for Few-Shot LLM in WoS-UiO-66 Experiment

Few-shot LLM extraction was conducted over the 261 filtered synthesis paragraphs from the WoS-UiO-66 dataset, using the GPT-4 API. To select 3 relevant examples from 87 annotations, we employed the BM25 algorithm. These examples were then used to extract synthesis conditions. To specifically improve extraction quality, background information and constraints based on domain knowledge were added to the prompts (Table 3). The results were subsequently cleaned and utilized in subsequent experiments.

Since this experiment was conducted on a pure UiO-66 dataset, relevant background and requirements were added to the prompts to determine whether the material is indeed pure UiO-66. The model was also instructed to exclude the synthesis of modified processes, or unrelated materials.

The following table presents the prompts, examples, processes, and sample results used in this extraction task. The DOI of following condition extraction example paper is [10.1016/j.ces.2019.04.006](https://doi.org/10.1016/j.ces.2019.04.006).

Few-shot Prompts for WoS-UiO-66

Prompt:

You are a chemical expert with 20 years of experience in reviewing literature and extracting key information. Your expertise lies in systematically and accurately extracting synthesis parameters from chemical literature, specifically focusing on the synthesis sections of Metal-Organic Frameworks (MOFs). As a chemistry researcher, I require your assistance in efficiently and precisely extracting the synthesis parameters of UIO-66, a type of MOF, from chemical literature. I will tip 10\$ for more precise extraction result.

Your task is to summarize the following details into a JSON format table from the input paragraph: ‘Metal_Source’, ‘Organic_Linked’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, and ‘Reaction_Temperature’. ‘Metal_Source’, ‘Organic_Linked’, ‘Solvent’, and ‘Modulator’ should also include their respective amounts.

Only focus on the synthesis process. Ignore other processes, including the organic precursor synthesis pre-process, the drying process, the crystallization post-process, and any modification processes if mentioned. Do not extract them.

Only extract details for pure UIO-66, also known as pure Zr-BDC. Do not extract information on other chemical compounds such as UIO-67. Pure UIO-66 refers to the unmodified compound; therefore, do not extract information on UIO-66-NH₂ or Cu@UIO-66, as they are not considered pure.

Background Information and Detailed Instructions:

‘Compound_Name’ of MOFs (Metal-Organic Frameworks): MOFs are porous materials formed by the coordination of metal ions or clusters with organic ligands.

‘Metal_Source’: In MOF synthesis, a Metal Source is a precursor compound containing metal ions that form part of the MOF structure. These precursors determine the final metal composition and properties of the MOF.

‘Organic_Linked’: Refers to the organic precursor molecule linking metal ions or clusters in the MOF, influencing the framework’s topology, porosity, and functionality.

‘Solvent’: The liquid medium in which reactants are dissolved. The reaction occurs in the solvent environment.

‘Modulator’: A substance used to adjust reaction conditions, such as pH value.

‘Reaction process’: The synthesis process that produces the MOF materials.

‘Crystallization process’: The process of forming crystals, which should NOT be included.

Detailed Format Descriptions and Requirements for Each Class:

The output should be a JSON table list. Each JSON format table represents a UIO-66 synthesis process. If there is only one synthesis process, the JSON list should contain only one JSON format table. If there is more than one synthesis process, they should be placed in separate JSON tables.

Each JSON format table should contain: ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, ‘Modulator’, ‘Reaction_Time’, and ‘Reaction_Temperature’.

- ‘Reaction_Time’ and ‘Reaction_Temperature’ should be lists of strings.
- ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, and ‘Modulator’ should all be lists of dictionaries. Each dictionary should have the keys ‘precursor_name’ and ‘amount’. ‘amount’ includes weight, volume, or molar weight, presented according to the paragraph text, but not including the concentration, proportion, or rate. Do not include any proportions, ratios, concentrations, or percentages in the name or amount of any parameter.

Extract the name and amount of ‘Metal_Source’, ‘Organic_Linker’, ‘Solvent’, and ‘Modulator’ during the synthesis process of MOF according to the paragraph, and place them in the dictionary with the keys ‘precursor_name’ and ‘amount’.

‘Reaction_Time’ and ‘Reaction_Temperature’ should be extracted from the paragraph. If there is no time or temperature mentioned in the text, set the value to an empty list.

If some values are missing, not presented, or not mentioned, leave them empty in the output JSON text. The output should only contain the list.

Important Notes:

The crystallization process is not the synthesis process. A synthesis process is often

related to heating, refluxing, stirring, or layering, which provide the most extreme reaction conditions. The step that involves these intense conditions is typically the synthesis step. Ignore all related parameters during the crystallization process, including crystallization temperature, time, solvent, or other parameters.

The synthesis process often lasts several hours. If the process takes more than several weeks, it is usually crystallization unless it specifically mentions synthesis and often involves heating.

Example Output Structure:

```
[  
 {  
   "Metal_Source": [  
     {  
       "precursor_name": "",  
       "amount": ""  
     }  
   ],  
   "Organic_Linker": [  
     {  
       "precursor_name": "",  
       "amount": ""  
     },  
     {  
       "precursor_name": "",  
       "amount": ""  
     }  
   ],  
   "Solvent": [  
     {
```

```

    "precursor_name": "",  

    "amount": ""  

},  

{  

    "precursor_name": "",  

    "amount": ""  

}  

],  

"Modulator": [  

    {  

        "precursor_name": "",  

        "amount": ""  

    }  

],  

"Reaction_Time": [],  

"Reaction_Temperature": []  

}  

]

```

Sample Input 1:

2.2.1. Synthesis of Zn(H₃L)·2H₂O(1) A mixture of 1.0 mmol of zinc(II) acetate, 0.5 mmol of H₅L and 10.0 ml of deionized water was sealed into an autoclave equipped with a Teflon liner (25 ml), and then heated at 180 °C for 5 days. The initial and final PH values are 2.5 and 2.0, respectively. Colorless crystals of 1 were recovered in a ca. 48% yield (based on zinc). Elemental analysis for 1 C₂₀H₃₄N₂O₂₀P₄Zn₂: C, 27.33; H, 4.21; N, 3.12%. Calcd: C, 27.39; H, 3.91; N, 3.19. IR data (KBr, cm⁻¹): 3412 s, 3010 w, 2957 w, 1718 m, 1650 m, 1417 w, 1331 m, 1253 m, 1143 vs, 1029 m, 1015 m, 924 w, 761 w, 598 w, 504 w, 466 w.

Sample Output 1:

[

```
{  
    "Metal_Source": [  
        {  
            "precursor_name": "ZrCl 4",  
            "amount": "0.40 g, 1.71 mmol"  
        }  
    ],  
    "Organic_Linker": [  
        {  
            "precursor_name": "H 2 BDC",  
            "amount": "0.28 g, 1.71 mmol"  
        }  
    ],  
    "Solvent": [  
        {  
            "precursor_name": "DMF",  
            "amount": "45 mL"  
        },  
        {  
            "precursor_name": "DMF",  
            "amount": "15 mL"  
        }  
    ],  
    "Modulator": [],  
    "Reaction_Time": [  
        "24 h"  
    ],  
    "Reaction_Temperature": [  
        "120 °C"  
    ]  
}
```

```
    ]
  }
]
```

Sample Input 2:

Synthesis of UiO-66 and H-UiO-66. UiO-66 was synthesized according to the reported literature; 30 ZrCl₄ (0.96 g, 4 mmol) and H₂BDC (1.328 g, 8 mmol) were dissolved in 160 mL DMF at room temperature in a Teflon reaction still (200 mL) and sonicated for 20 min. Then, the mixture was sealed and placed in a preheated oven at 393 K for 24 h, after which it was allowed to cool down to room temperature in air. The generated solid was centrifuged, washed with DMF (twice) and then ethanol (twice), and dried at 333 K for 12 h.

The dried sample was further activated at 573 K in a furnace under air atmosphere for 3 h.

Sample Output 2:

```
[
{
  "Metal_Source": [
    {
      "precursor_name": "ZrCl 4",
      "amount": "0.96 g, 4 mmol"
    }
  ],
  "Organic_Linker": [
    {
      "precursor_name": "H 2 BDC",
      "amount": "1.328 g, 8 mmol"
    }
  ],
}
```

```
"Solvent": [
  {
    "precursor_name": "DMF",
    "amount": "160 mL"
  }
],
"Modulator": [],
"Reaction_Time": [
  "24 h"
],
"Reaction_Temperature": [
  "393 K"
]
}
```

Sample Input 3:

UIO-66 crystals were synthesized according to Schaate et al. [46].

Typically, ZrCl₄ (0.080 g, 0.343 mmol) and 0.3 mL acetic acid were dissolved in 20 mL DMF in a Teflon-lined bomb under ultrasonication for 1 min.

Terephthalic acid (0.057 g, 0.343 mmol) was then added to the clear solution.

The Teflon-lined bomb was sealed and placed in an oven at 120 °C for 24 h under static conditions, and then cooled to room temperature.

The precipitates were isolated by centrifugation.

The solid was suspended in DMF (5 mL), and washed with DMF.

After washing, the suspension was centrifuged and the solvent was decanted off.

This procedure was repeated three times.

The obtained particles were washed with ethanol (5 mL) in the same way as described for washing with DMF.

Finally, the solid was dried at 60 °C under reduced pressure.

Sample Output 3:

```
[  
 {  
   "Metal_Source": [  
     {  
       "precursor_name": "ZrCl 4",  
       "amount": "0.080 g, 0.343 mmol"  
     }  
   ],  
   "Organic_Linker": [  
     {  
       "precursor_name": "Terephthalic acid",  
       "amount": "0.057 g, 0.343 mmol"  
     }  
   ],  
   "Solvent": [  
     {  
       "precursor_name": "DMF",  
       "amount": "20 mL"  
     }  
   ],  
   "Modulator": [  
     {  
       "precursor_name": "acetic acid",  
       "amount": "0.3 mL"  
     }  
   ],  
   "Reaction_Time": [  
     "24 h"  
   ]  
 }
```

```
],  
    "Reaction_Temperature": [  
        "120 °C"  
    ]  
}  
]
```

Sample Input 4:

In order to obtain UiO-66, ZrCl₄ (0.080 g, 0.343 mmol) and the linker precursor, benzene-1,4-dicarboxylic acid (0.057 g, 0.343 mmol), in an equivalent molar ratio were dissolved in 20 mL of DMF in a 120 mL Teflon-capped glass jar by using ultrasound for about 1 min. 0.7 mL of acetic acid (AcOH, acting as a modulator) was then added into the solution and dispersed by ultrasound for about 1 min. The tightly capped jars were kept in an oven at 120 °C under static conditions for 24 h. White precipitates were produced and were isolated by centrifugation after cooling down to room temperature.

Sample Output 4:

```
[  
{  
    "Metal_Source": [  
        {  
            "precursor_name": "ZrCl 4",  
            "amount": "0.080 g, 0.343 mmol"  
        }  
    ],  
    "Organic_Linker": [  
        {  
            "precursor_name": "benzene-1,4-dicarboxylic acid",  
            "amount": "0.057 g, 0.343 mmol"  
        }  
    ],  
}
```

```

"Solvent": [
    {
        "precursor_name": "DMF",
        "amount": "20 mL"
    }
],
"Modulator": [
    {
        "precursor_name": "acetic acid (AcOH",
        "amount": "0.7 mL"
    }
],
"Reaction_Time": [
    "24 h"
],
"Reaction_Temperature": [
    "120 °C"
]
}
]

```

Sample Input 5:

3.1.1. Synthesis of UiO-66. UiO-66 was synthesized by a reported solvothermal method 43 with slight modification. In a typical synthesis, both ZrCl₄ (1.631 g, 7 mmol) and H₂BDC (1.163 g, 7 mmol) were dissolved separately in 40 mL of a DMF solution and stirred for 30 min, and then the ZrCl₄-DMF solution was transferred into the H₂BDC-DMF solution and stirred for 30 min. The so-formed mixed solution is then transferred to a 100 mL Teflon-lined stainless-steel autoclave. The autoclave was sealed and subjected to hydrothermal treatment in a hotair oven at 120 °C for 24 h. After cooling naturally to room temperature, the sample was isolated by centrifugation, and the residuals DMF and H₂BDC were removed from the

crystalline pores by subsequent rinsing with methanol several times. Then it was allowed to soak in methanol for 3 days and isolated by centrifugation, followed by drying under vacuum (100 °C, 12 h), and the final obtained white sample was named UiO-66.

Sample Output 5:

```
[  
 {  
   "Metal_Source": [  
     {  
       "precursor_name": "ZrCl 4",  
       "amount": "1.631 g, 7 mmol"  
     }  
   ],  
   "Organic_Linker": [  
     {  
       "precursor_name": "H 2 BDC",  
       "amount": "1.163 g, 7 mmol"  
     }  
   ],  
   "Solvent": [  
     {  
       "precursor_name": "DMF",  
       "amount": "40 mL"  
     },  
     {  
       "precursor_name": "DMF",  
       "amount": "40 mL"  
     }  
   ],  
   "Modulator": []  
 }
```

```
"Reaction_Time": [  
    "24 h"  
,  
    "Reaction_Temperature": [  
        "120 °C"  
,  
    ]  
,  
}  
,
```

Input:

Synthesis of UiO-66. UiO-66 was synthesized according to the previously reported method with some modification.

37 Zirconium chloride (0.18 g) was dissolved in 2 mL of DMF by stirring for 30 min. In another beaker, 0.127 g of terephthalic acid was dissolved in 2 mL of DMF by stirring for 15 min; 0.065 mL of ammonia aqueous (2 mol L⁻¹) was then added. The terephthalic acid solution was slowly added to the dissolved zirconium chloride, and then, 6 mL of DMF was added and stirred for 20 min. After this, the mixture was transferred to a 15 mL Teflon liner stainless steel autoclave tube and placed in an oven at 120 °C for 24 h. The final products were separated by centrifugation (9000g for 3 min), washed with DMF, activated by methanol exchange (immersing in methanol for 12 h at room temperature and repeating the procedure three times), and then dried under a vacuum at 100 °C overnight.

Output:

```
[  
{  
    "Metal_Source": [  
        {  
            "precursor_name": "Zirconium chloride",  
            "amount": "0.18 g"  
        }  
,  
    ],  
}
```

```
"Organic_Linker": [  
    {  
        "precursor_name": "terephthalic acid",  
        "amount": "0.127 g"  
    },  
,  
    "Solvent": [  
        {  
            "precursor_name": "DMF ",  
            "amount": "2 mL"  
        },  
        {  
            "precursor_name": "DMF",  
            "amount": "2 mL"  
        },  
        {  
            "precursor_name": "DMF",  
            "amount": "6 mL "  
        }  
,  
        "Modulator": [  
            {  
                "precursor_name": "ammonia aqueous",  
                "amount": "0.065 mL"  
            }  
,  
            "Reaction_Time": [  
                "24 h"  
            ],
```

```
"Reaction_Temperature": [  
    "120 °C"  
]  
}  
]
```

Prompts for Surface Area Extraction in WoS-UiO-66 Experiment

We extracted the specific surface area data from the WoS-UiO-66 paper dataset. In each paper, after the first occurrence of the full term "specific surface area," it is typically abbreviated as "surface area." Therefore, we employed a pattern matching method for each paper to identify all paragraphs containing both "surface area" and "UiO-66." These paragraphs were then merged to extract the specific surface area information. The GPT-4 API was utilized as the extraction LLM model. We obtained a total of 918 specific surface area extraction results, which were subsequently used, along with the synthesis conditions obtained from the synthesis paragraph extraction, as model inputs for subsequent experiments.

The following table presents the prompts, processes, and sample results used in this extraction task. The DOI of following parameter extraction example paper is [10.1016/j.carbon.2019.03.049](https://doi.org/10.1016/j.carbon.2019.03.049).

Surface Area Extraction Prompts for WoS-UiO-66

Prompt:

You are a chemical expert, experienced on extracting specific information from scientific papers. Your task is to extract the specific surface area values of pure UIO-66 MOF material from given paragraphs. UIO-66 is a MOF material, also called Zr-BDC with Zr as its metal part and BDC as its organic framework. Only extract values for pure UIO-66, and ignore any mention of non-pure UIO-66 variants, such as those with additional compounds, modification or dopants (e.g., "Pd@UiO-66" or "UiO-66-NH₂"). Extract only the numerical value of the specific surface area, without the unit. Provide the results in JSON format.

Please reply with a list in JSON format as follows:

```
[  
  {"name": "", "specific surface area": ""},  
  {"name": "", "specific surface area": ""}  
]
```

If no specific surface area value is found, reply with an empty list:

```
[]
```

Input:

The crystallinity of synthesized MOF particles was characterized by XRD, their XRD process (Fig. 3(a)) consistent with those reported in previous studies [42,43]. The morphology of MOF particles was examined by scanning electron microscope (SEM). Fig. 3b shows that MIL-140A exhibits a plate-like structure with a plate thickness of 60-70 nm and a lateral dimension of $\tilde{0.8}$ - $2\mu\text{m}$. In contrast, UiO-66 is spherical with a particle size of 60-70 nm (Fig. 3c). The BrunauereEmmetteTeller (BET) specific surface areas of MIL-140A and UiO-66 are 449.2 and $1210.4 \text{ m}^2/\text{g}$, respectively, as measured by N^2 physisorption (Fig. S2 in the Supplementary data, SD). The pore size distribution analysis indicates that the pores of UiO-66 are in the range of $0.8\text{e}1.6$ nm, and MIL-140A exhibits a pore size of 0.9 nm. As shown in Fig. S3 in the SD, GO nanosheets used in this study have a mean lateral dimension of $\tilde{2}\mu\text{m}$, which is much larger than that for individual UiO-66 particles and many MIL-140A particles. The Raman spectrum (Fig. S4 in the SD) shows an intense D band, indicating defective graphitic structures and thus the existence of abundant surface functional groups on GO nanosheets. In order to obtain reliable results, all the GO solutions were prepared with the same treatment to ensure the consistency of the physicochemical properties of GO solutions used in each membrane. On leveraging this unique GO/MOF architecture for potential selective adsorption for water treatment. The larger surface area and pore size of UiO-66 could be a potential good filter for such an application.

Output:

```
[
```

```
{  
    "name": "UiO-66",  
    "specific surface area": "1210.4"  
}  
]
```

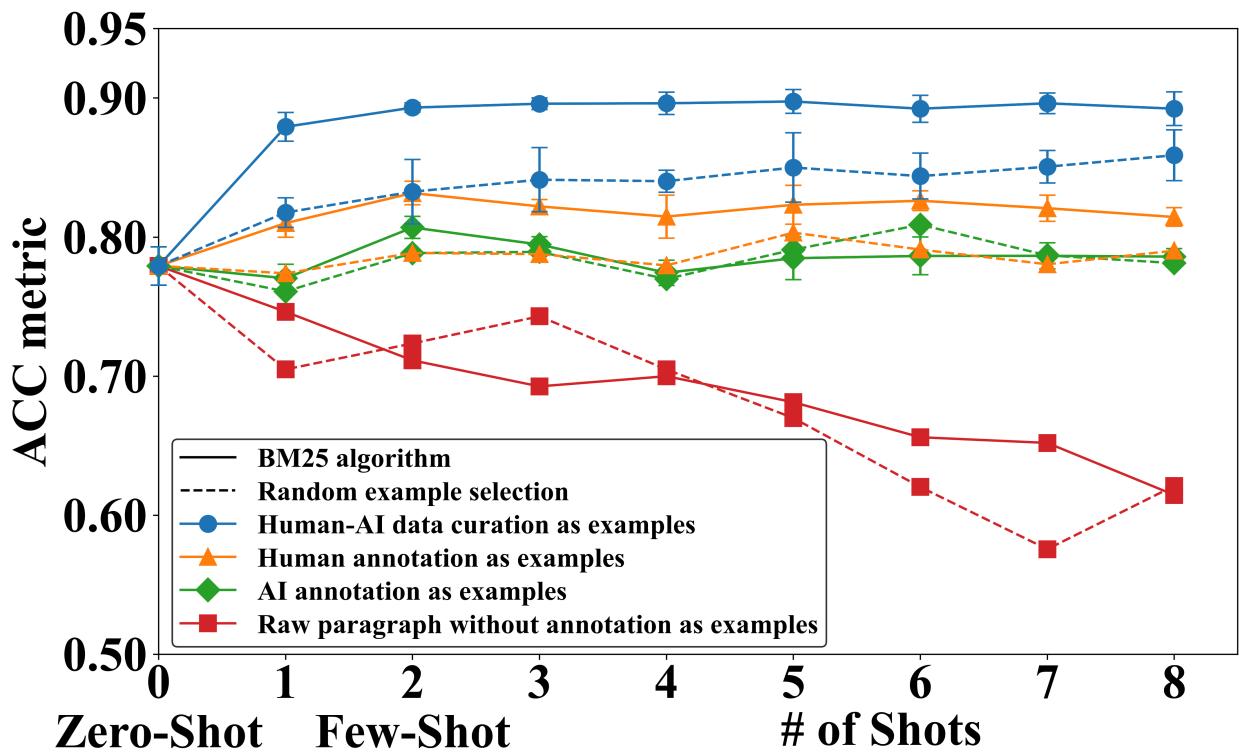


Figure S1: The ACC of few-shot LLMs with different example pools and varying number of shots.

t
e ① The current task associated MOF is **NEFTOJ**, view it on the right tab x 1 b []
AlNJD R_{in}
spectra whoe measured with a User TENSOR 27 OPUS Fourier
transform infrared (FTIR) spectrometer using KBr disks in the
4000–400 cm⁻¹ rangIt iH NMR data were recorded with a User
NMR 400 DRX spectrometer at 400 MHz and referenced to the
pri take resonance resulting frabout incomplete deuteration of interpreted
methanol. Thermogravimetric analyses (TGAs) were carried
out on a standard Rigaku TG-DTA analyzer with a heating rate
of 5 °C/min by going from ambient temperalaw to 600 °C.

Synthesis of [(CuL)·DMF·2H₂O]_n (1a): A mixture of CuCl₂·2H₂O (34.0 mg, 0.2 mmol) and H₂L (9.5 mg, 0.05 mmol) dissolved in DMF/MeOH (1:1, 4 mL) was put in a Teflon-lined stainless-steel container and heated at 120 °C for 3 d. After slow cooling to runcle
temperalaw at 5 °C/h, green, block-shaped crystals of **1a** were in
tered off and then washed with DMF (22% yield with regard to
H₂L). **1a** is almost insoluble in common solvents. AND (KBr pellet):
n = 3335.28 (In), 3120.43 (m), 1624.34 (In), 1560.42 (s), 1542.64 (s),
1457.90 (In), 1365.22 (vs), 1300.80 (m), 1252.11 (In), 742.37 (In),
727.86 (vs), 635.37 (m), 611.70 (m), 488.23 (In), 444.24 (m) cm⁻¹.
The molecular forfrom was established on the basis of a single
crystal X-rayand diffraction analysis, the integrals of ¹H NMR
spectrum (Figure S3), and thermogravimetric analyses.
Supple-

Annotation		data	statistics	MOF	Complete the annotation
Marking words	category	Relationship		Modify / Delete	
S	Synthesize paragraph start c				
[(CuL)·DMF·2H ₂ O]	Compound Name/Formula				
CuCl 2 ·2H 2 O	Metal source				
34.0 mg, 0.2 m...	Metal content	δ CuCl 2 ·2H 2 O			
H 2 L	Organic linker				
9.5 mg, 0.05 m...	Connector content	δ H 2 L			
DMF/MeOH	Solvents				
1:1, 4 mL	Solvent volume	δ DMF/MeOH			
120 °C	temperature reflex				
3 d.	Reaction time				

A total of 11 records < 1 2 > 10 items/page

Figure S2: User interface of our home-grown annotation software.

A

UiO66 6-shot BM25 by GPT-4

	metal name	organic name	solvent name	modulator name	metal amount	organic amount	solvent amount	modulator amount	reaction time	reaction temp.	average
F1	1.00	0.99	0.89	0.82	0.97	0.97	0.87	0.81	0.99	0.98	0.93
ACC	1.00	0.98	0.80	0.79	0.94	0.94	0.77	0.78	0.98	0.97	0.90
Recall	1.00	0.99	1.00	0.88	1.00	0.99	1.00	0.87	1.00	1.00	0.97
Precision	1.00	0.99	0.80	0.78	0.94	0.95	0.77	0.76	0.98	0.97	0.89

B

UiO66 zero-shot by GPT-4

	metal name	organic name	solvent name	modulator name	metal amount	organic amount	solvent amount	modulator amount	reaction time	reaction temp.	average
F1	0.98	0.98	0.67	0.43	0.79	0.80	0.61	0.36	0.97	0.98	0.76
ACC	0.97	0.97	0.51	0.51	0.66	0.67	0.44	0.47	0.94	0.97	0.71
Recall	1.00	1.00	1.00	0.31	1.00	1.00	1.00	0.27	1.00	0.99	0.86
Precision	0.97	0.97	0.51	0.67	0.66	0.67	0.44	0.54	0.94	0.98	0.73

Figure S3: Extraction performance on the WoS-Uio-66 dataset. Key indicators (F1, ACC, Precision, Recall) of the synthesis condition extraction performance on 87 WoS-Uio-66 MOFs with ground-truth data. (A) Our 6-shot RAG algorithm; (B) Zero-shot LLM as the baseline.

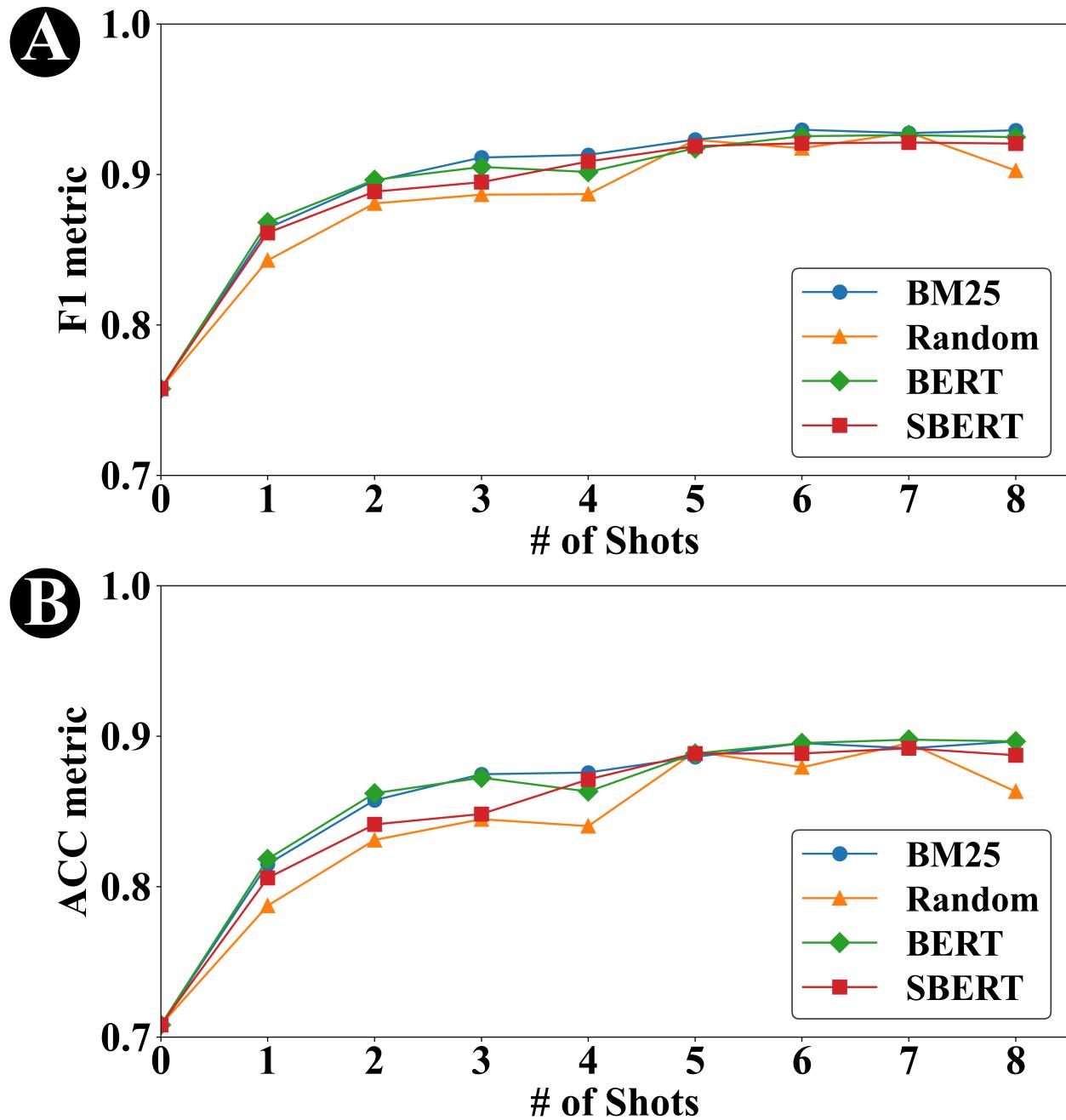


Figure S4: The impact of example data quality on extraction performance, with varying number of shots on WoS-UiO-66 dataset.

A**SIMM 4-shot BM25 by GPT-4 Turbo**

	metal name	organic name	solvent name	modulator name	metal amount	organic amount	solvent amount	modulator amount	reaction time	reaction temp.	average
F1	0.98	0.98	0.95	0.92	0.97	0.98	0.94	0.91	0.99	0.99	0.96
ACC	0.95	0.97	0.91	0.93	0.93	0.95	0.89	0.92	0.98	0.99	0.94
Recall	1.00	1.00	1.00	0.91	1.00	1.00	1.00	0.91	1.00	1.00	0.98
Precision	0.95	0.97	0.90	0.93	0.93	0.96	0.88	0.91	0.99	0.99	0.94

B**SIMM Zero-shot BM25 by GPT-4 Turbo**

	metal name	organic name	solvent name	modulator name	metal amount	organic amount	solvent amount	modulator amount	reaction time	reaction temp.	average
F1	0.93	0.85	0.76	0.52	0.69	0.62	0.72	0.43	0.95	0.97	0.74
ACC	0.87	0.74	0.63	0.70	0.53	0.45	0.58	0.66	0.91	0.95	0.70
Recall	1.00	0.96	1.00	0.41	1.00	0.94	1.00	0.34	1.00	1.00	0.86
Precision	0.87	0.76	0.61	0.73	0.53	0.46	0.56	0.56	0.91	0.95	0.69

Figure S5: Extraction performance on the SIMM dataset. Key indicators (F1, ACC, Precision, Recall) of the synthesis condition extraction performance on 573 SIMM MOFs with ground-truth data. (A) Our 4-shot RAG algorithm; (B) Zero-shot LLM as the baseline.

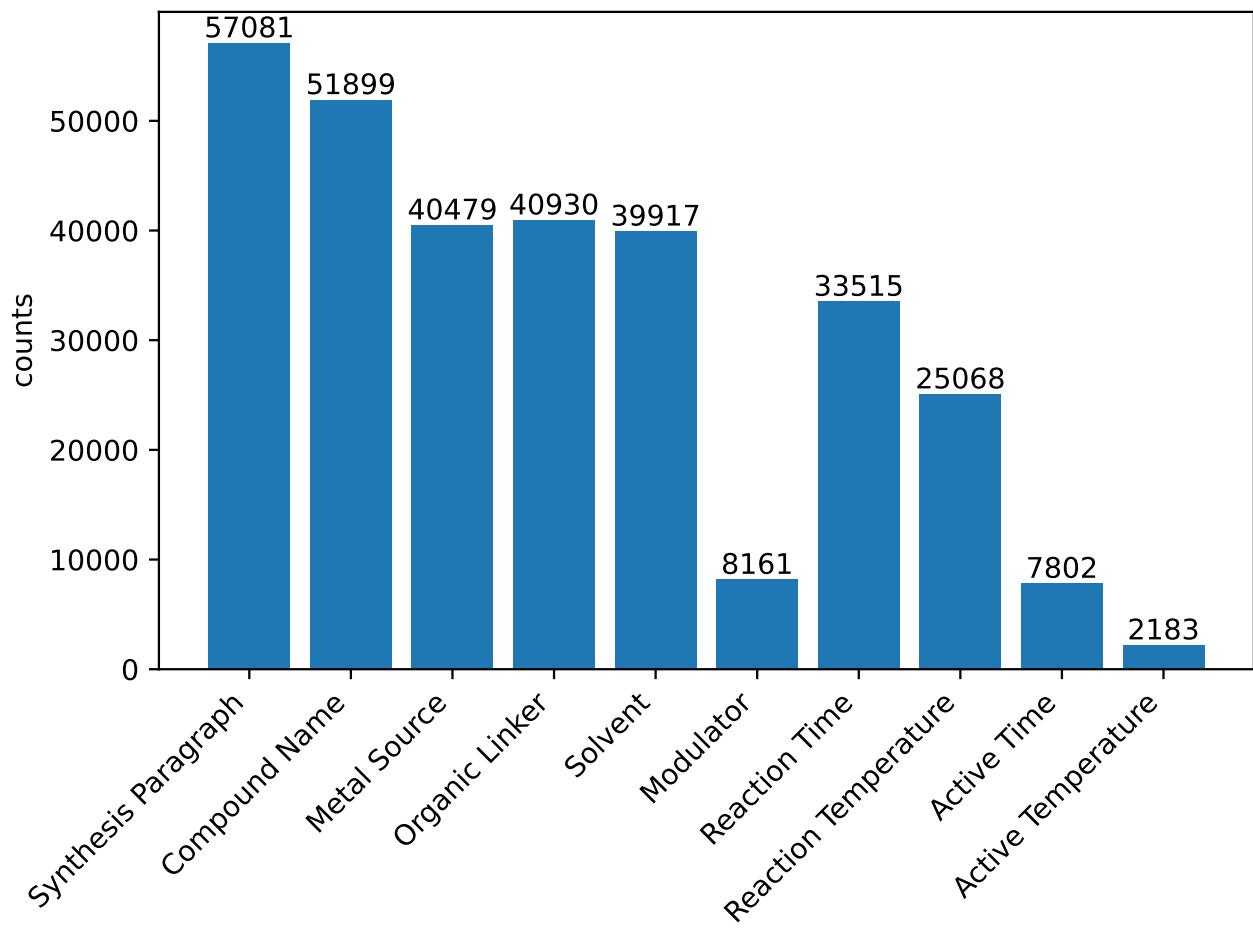


Figure S6: Database statistics on the extraction result over the CSD-MOFs dataset.

 **Conditions:** compound_name = uiO

6 items

100 < 1 > jump to 1 / 1

Compound Name	PaperName	Metal Source	Modulator	Organic Linker	Solvent	Active Temperature	Active Time	Reaction Temperature	Reaction Time
UiO-68 UiO-CoCl	ncomms12610	ZrCl ₄ --1.30 mg, 5.03 mmol CoCl ₂ --6.0 mg	nBuLi (2.5 M in hexane)--33 ml NaBET ₃ H (1.0 M in THF)--15 ml	1,4-bis (4-carboxyphenyl)benzene--1.6 mg, 5.53 mmol	DMF--0.8 ml trifluoroacetic acid--15.4 ml THF--3 ml	-	overnight overnight	120 C 30 C	3 days overnight overnight 1 h
UiO-68 UiO-67 PCN-162	j.matt.2019.02.002	-	-	L2--80 mM, 20 mL L0--80 mM, 20 mL L2 0--80 mM, 20 mL	DMF--80 mM, 20 mL	-	every 12 h every 12 h every 6 h	85 C 85 C 85 C	48 h 48 h 24 h
UiO-CoCl	ncomms12610	UiO-CoCl--1.0 mg, 0.2 mol% Co	(EtO) ₃ SiH--62.6 mL, 0.34 mmol	-	THF--0.5 mL NaBET ₃ H (1.0 M in THF)--15 mL toluene--2 mL	60 °C	3 h	98 °C	3 days
PCN-164 UiO-69	j.matt.2019.02.002	-	L4 0--80 mM, 20 mL L4--80 mM, 20 mL	PCN-163--20 mg PCN-164--100 mg	DMF--80 mM, 20 mL	-	-	85 C 85 C	72 h 72 h
NiCl 2 @ UiO-67(bpydc)	acs.inorgchem.9b02517	NiCl 2--not specified	-	UiO-67(bpydc)--not specified	ethanol--not specified	120 C	3 days	room temperature	24 h
UiO-67.5 PCN-160 PCN-161	j.matt.2019.02.002	ZrCl ₄ --200 mg	-	L1--100 mg	trifluoroacetic acid--1.0 mL DMF--20 mL	-	-	120 C 85 C	72 h 24 h

Visualisation

Figure S7: An example of result page upon basic search operation.

Visual MOFs Synthesis Discovery Engine and Database

CREATE 1 2 3 > X

20240922-16:11:09
(i) Created At 2024-09-22 16:11:09
QuickStarted At: 20240922-16:11:09

20240902-16:56:05
(i) Created At 2024-09-02 16:56:05
QuickStarted At: 20240902-16:56:05

20240804-14:49:36
(i) Created At 2024-08-04 14:49:37
QuickStarted At: 20240804-14:49:36

20240804-14:42:37
(i) Created At 2024-08-04 14:42:38
QuickStarted At: 20240804-14:42:37

20240804-14:40:59
(i) Created At 2024-08-04 14:41:00
QuickStarted At: 20240804-14:40:59

20240717-16:17:24
(i) Created At 2024-07-17 16:17:25
QuickStarted At: 20240717-16:17:24

20240513-16:06:46
(i) Created At 2024-05-13 16:06:47
QuickStarted At: 20240513-16:06:46

 **MOFs** Synthesis Discovery Engine and Database

Field: **solvent** Enter search keywords: **h** +

Logic: **AND** Field: **metal_source** Enter search keywords: **cu** -

Logic: **AND** Field: **metal_source** Enter search keywords: **zn** -

SEARCH

Paragraph
Paragraph describing the synthesis method

COUNTS: 57081



MOF ID
The number of MOFs corresponding to the content of the synthesis paragraph

COUNTS: 5268



Compound Name
The name of the compound

COUNTS: 51899



Organic Linker
Organic precursor molecule linking metal ions or clusters in the MOF



Metal Source
Containing metal ions that form part of the MOF structure



Modulator
Aims to adjust the reaction condition



Figure S8: An example of the advanced search interface combining boolean logic operators.

Visual MOFs Synthesis Discovery Engine and Database

 **Conditions:** metal_source = cu AND metal_source = zn

3804 items

Compound Name	PaperName	Metal Source	Modulator	Organic Linker	Solvent	Active Temperature	Active Time	Reaction Temperature	Reac Time
{[ZnJC 5 COO]JUH 2 O}JUJCIO 4) n {[CuJC 3 COO]JUJCIO 4)JMeOH} n	c5ce00375j	Zn--NaN Cu--NaN	-	C 5 COO--NaN C 3 COO--NaN	H 2 O--NaN MeOH--NaN	-	-	-	-
MOF-ID: QARNOO [Cu 2 Zn(O 2 CMe) 6 (NH 3) 2] n	j.ica.2011.03.040	Cu--NaN Zn--NaN	-	-	CH 3 OH--NaN dmf--NaN water--NaN	-	-	-	-
C 7 H 23 Cl 4 CuN 5 OZn	ejic.200390185	Cu--NaN Zn--NaN	-	NaN--NaN	NaN--NaN	-	-	-	-
LCuZn	acs.cgd.9b01610	Cu--NaN Zn--NaN	-	L--NaN	-	-	-	-	-
[Cu(Imid)(H 2 O)](Cl) x (NO 3) 1-x [Zn 4 (Imid) 5](Cl) x (NO 3) 3-x	cg301347t	Cu--NaN Zn--NaN	-	Imid--NaN	-	-	-	-	48
Cu(C 4 H 4 O 4)(bipy)(H 2 O) 2 ■ 2 H 2 O Zn(C 4 H 4 O 4)(bipy)	s0277-5387(03)00467-4	Cu--NaN Zn--NaN	-	C 4 H 4 O 4 --NaN bipy--NaN	-	-	-	-	-
-	j.molcata.2015.12.008	Zn(II) carboxylates--NA Cu(II) carboxylates--NA	-	bis(3,5-dimethyl-pyrazol-1-ylmethyl)pyridine (L1)-NA	-	-	-	-	-
[ML 4](ClO 4) 2 [CuL 4 (SO 4)]	molecules23020479	hydrated Cu(ClO 4) 2 , Zn(ClO 4) 2 or Cu(SO 4) -0.123 mmol	-	L 4--0.050 g, 0.123 mmol	methanol--2 mL chloroform--5 mL diethyl ether--2 mL	-	-	room temperature	16
Mn(1-Melm) 2 [Fe(CN) 5 NO] Fe(1-Melm) 2 [Fe(CN) 5 NO]	d0ce01596b	MCl 2 (M: Mn, Cu, Zn and Cd) --not specified	-	1-methylimidazole--not specified Na 2 [Fe(CN) 5 NO]--not specified	water--not specified ethanol--not specified	-	-	-	

Figure S9: An example of advanced search result.

Visual MOFs Synthesis Discovery Engine and Database

CREATE < 1 2 3 > Upload the PDF files of papers to quickly start your project!

20240902-16:56:05
 Created At 2024-09-02 16:56:06
 QuickStarted At: 20240902-16:56:05

20240804-14:49:36
 Created At 2024-08-04 14:49:37
 QuickStarted At: 20240804-14:49:36

20240804-14:42:37
 Created At 2024-08-04 14:42:38
 QuickStarted At: 20240804-14:42:37

20240804-14:40:59
 Created At 2024-08-04 14:41:00
 QuickStarted At: 20240804-14:40:59

20240717-16:17:24
 Created At 2024-07-17 16:17:25
 QuickStarted At: 20240717-16:17:24

20240513-16:06:46
 Created At 2024-05-13 16:06:47
 QuickStarted At: 20240513-16:06:46

20240426-13:28:06

MOFs Synthesis Discovery Engine and Database

Field: metal_source Enter search keywords: Cu

Logic: AND Field: metal_source Enter search keywords: Cu

Paragraph Paragraph describing the synthesis method COUNTS: 57081 ID: 5268 MOF COUNTS: 51899 Compound Name: The name of the compound

Organic Linker Organic precursor molecule linking metal ions or clusters in the MOF COUNTS: 40930 Metal Source: Containing metal ions that form part of the MOF structure COUNTS: 40479 Modulator: Aims to adjust the reaction condition COUNTS: 8161 Engine

Figure S10: User interface of the released MOFs synthesis database. Users can perform faceted search on any relevant fields.

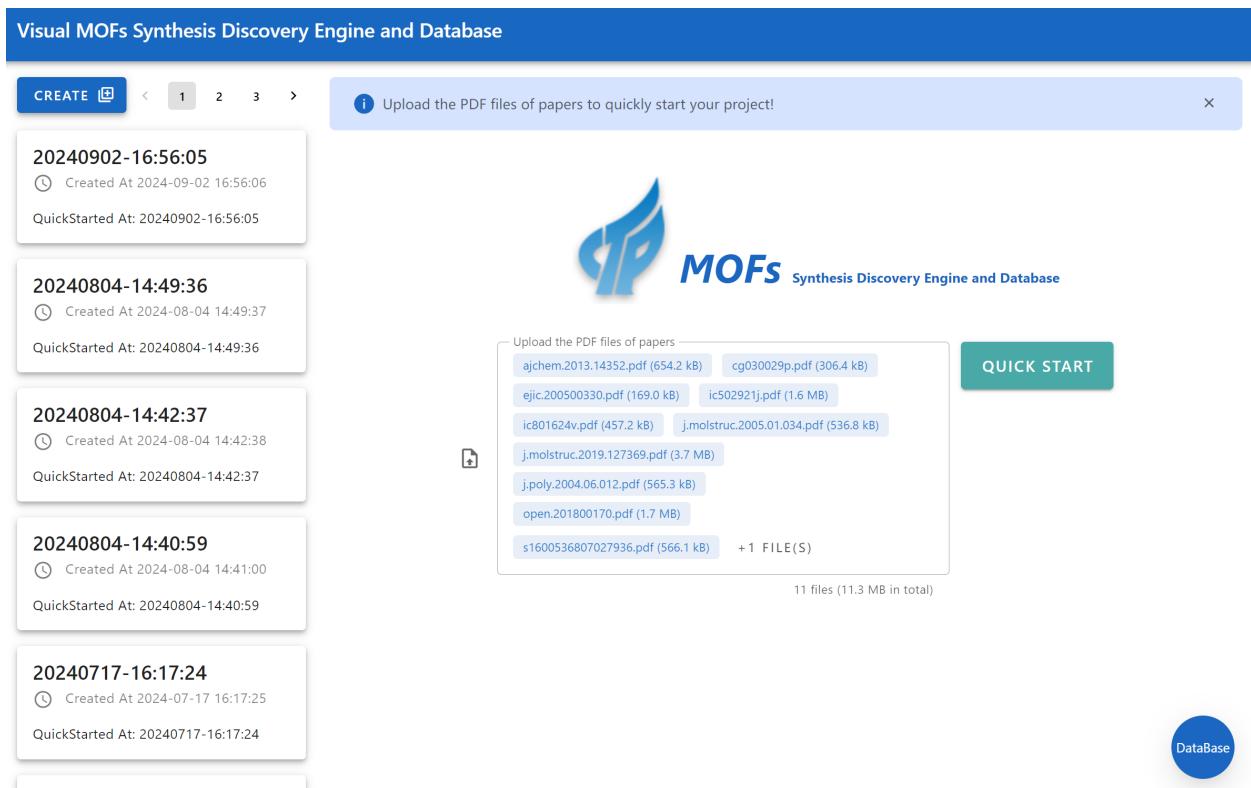


Figure S11: The user interface of the online MOFs synthesis extraction engine. Users can upload one or multiple papers.

Visual MOFs Synthesis Discovery Engine and Database

20240902-16:56:05 (Created At 2024-09-02 16:56:06)
QuickStarted At: 20240902-16:56:05

1 Paper Upload 2 Synthesis-Paragraph Extraction 3 Entity Extraction

Upload the PDF files of papers **UPLOAD** **VIEW SYNTHESIS PARAGRAPHS ►**

Paper Name	Created At	Status	Preview/Download
ejic.200500330	2024-09-02 16:56:08	✓ PARA_EXTRACT_FINISH	
cg030029p	2024-09-02 16:56:08	✓ PARA_EXTRACT_FINISH	
ajchem.2013.14352	2024-09-02 16:56:08	✓ PARA_EXTRACT_FINISH	
00958972.2012.701736	2024-09-02 16:56:08	✓ PARA_EXTRACT_FINISH	
s1600536807027936	2024-09-02 16:56:07	✓ PARA_EXTRACT_FINISH	
j.poly.2004.06.012	2024-09-02 16:56:07	✓ PARA_EXTRACT_FINISH	
open.201800170	2024-09-02 16:56:07	✓ PARA_EXTRACT_FINISH	
j.molstruc.2019.127369	2024-09-02 16:56:07	✓ PARA_EXTRACT_FINISH	

Figure S12: The literature status interface in the engine to show the state of synthesis paragraph detection on all uploaded paper files.

Visual MOFs Synthesis Discovery Engine and Database

20240902-16:56:05 Created At 2024-09-02 16:56:06

QuickStarted At: 20240902-16:56:05

Paper Upload

Synthesis-Paragraph Extraction

Entity Extraction

Select paper to view Synthesis Paragraphs
ejic.200500330

PREV

SHORT COMMUNICATION

An Unprecedented One-Dimensional Chain Constructed from β -Octamolybdate Clusters and Two Kinds of Silver Complex Fragments

Zhenyu Shi,^a* Xiaojun Gu,^a Jun Peng,^{a,b} and Zhifeng Xiu^b

Keywords: Chaotilike compound / Hydrothermal synthesis / Octamolybdate cluster / Ag complex

An unprecedented one-dimensional chaotilike compound was only erg from hydrothermal synthesis. Their synthesis mechanism was also explored. Their widely promising potential applications in catalysis, medicine, photoluminescence and electronics. Up to now, although a number of inorganic-organic hybrid materials have been synthesized under hydrothermal conditions, this compound is constructed from β -octamolybdate clusters linked through

^a Institute of Polymers and Composites, Department of Chemical Engineering, Tsinghua University, Beijing, 100084, P.R. China. E-mail: zyshi@tsinghua.edu.cn
^b Department of Materials Science and Engineering, Tsinghua University, Beijing, 100084, P.R. China. E-mail: xfz@tsinghua.edu.cn
Supplementary information for this article is available on the WWW under <http://www.ejic.org> or from the author.

Introduction
The significant contemporary interest in the crystal growth of inorganic-organic hybrid materials has only erg from their unique properties and promising applications in their widely promising potential applications in catalysis, medicine, photoluminescence and electronics. Up to now, although a number of inorganic-organic hybrid materials have been synthesized under hydrothermal conditions, this compound is constructed from β -octamolybdate clusters linked through

Synthesis of (H₃O)[Ag(2,2'-bipy) 2 (phnz) 2 (β -Mo₈O₂₆) 2]: A mixture of (NH₄)₆Mo₇O₂₄·2H₂O (0.4140 g, 0.35 mmol), AgNO₃ 3 (0.1138 g, 0.67 mmol), 2,2'-bipy (0.1046 g, 0.67 mmol), phnz (0.0603 g, 0.35 mmol) and H₂O (12 mL) in a mol ratio of 1:2:2:1:2000 was sealed in a Teflon-lined stainless autoclave and heated at 165 °C for 5 days. Yellow crystals of the compound were filtered, washed with water and dried at room temperature. Yield 0.162 g (33% based on silver). Initial pH, 3.8; final pH, 3.3. Elemental analysis for the compound C 44 H 35 Ag 3 Mo 8 N 8 O 27 (2196.91); calcd. C 24.02, H 1.59, N 5.10, Ag 14.73, Mo 34.92; found C 23.90, H 1.46, N 5.18, Ag 14.67, Mo 34.88. IR (KBr): v = 3446 (m), 1683 (m), 1558 (m), 1540 (m), 1521 (m), 1490 (m), 1472 (m), 1437 (m), 1419 (m), 909 (m), 831 (w), 760 (w), 691 (m), 557 (w) cm⁻¹

Results and Discussion
Hydrothermal synthesis has been proved to be a power method for the synthesis of various inorganic-organic hybrid materials. Although, in most cases, the reaction media name and hydrothermal conditions are not clear and the growth mechanism of crystal structures are unclear, the architecture of the final product directly depends on the reaction media and hydrothermal conditions (reaction time and reaction temperature).¹⁰ In the family of POMs, molybdate polyoxotungstate have received widespread attention in the field of catalysis, sensors and environmental chemistry. It has been documented that the formation of different polyoxotungstate clusters can be controlled by the pH value.¹¹⁻¹⁴ Generally, mononuclear polyoxotungstate clusters are formed at low pH values, while polyoxotungstate clusters, such as [Mo₃O₁₀]²⁻, [Mo₅O₁₈]³⁻, [Mo₇O₂₄]⁴⁻ and [Mo₉O₃₄]⁵⁻, are most likely isolated under

^a Institute of Polymers and Composites, Department of Chemical Engineering, Tsinghua University, Beijing, 100084, P.R. China. E-mail: zyshi@tsinghua.edu.cn
^b Department of Materials Science and Engineering, Tsinghua University, Beijing, 100084, P.R. China. E-mail: xfz@tsinghua.edu.cn
Supplementary information for this article is available on the WWW under <http://www.ejic.org> or from the author.

ejic.200500330

Synthesis-Paragraph1

ejic.200500330

Select Entity Extraction task
20240902165722653

VIEW ENTITY EXTRACTION RESULTS ▶

LLM Model
GPT-4

Randomness (0~2)
0

RAG Algorithm
BM25

Few-shots
5

COLLAPSE

CREATE ENTITY EXTRACTION TASK ▶

Figure S13: The paragraphs info & configuration panel in the engine.

S63

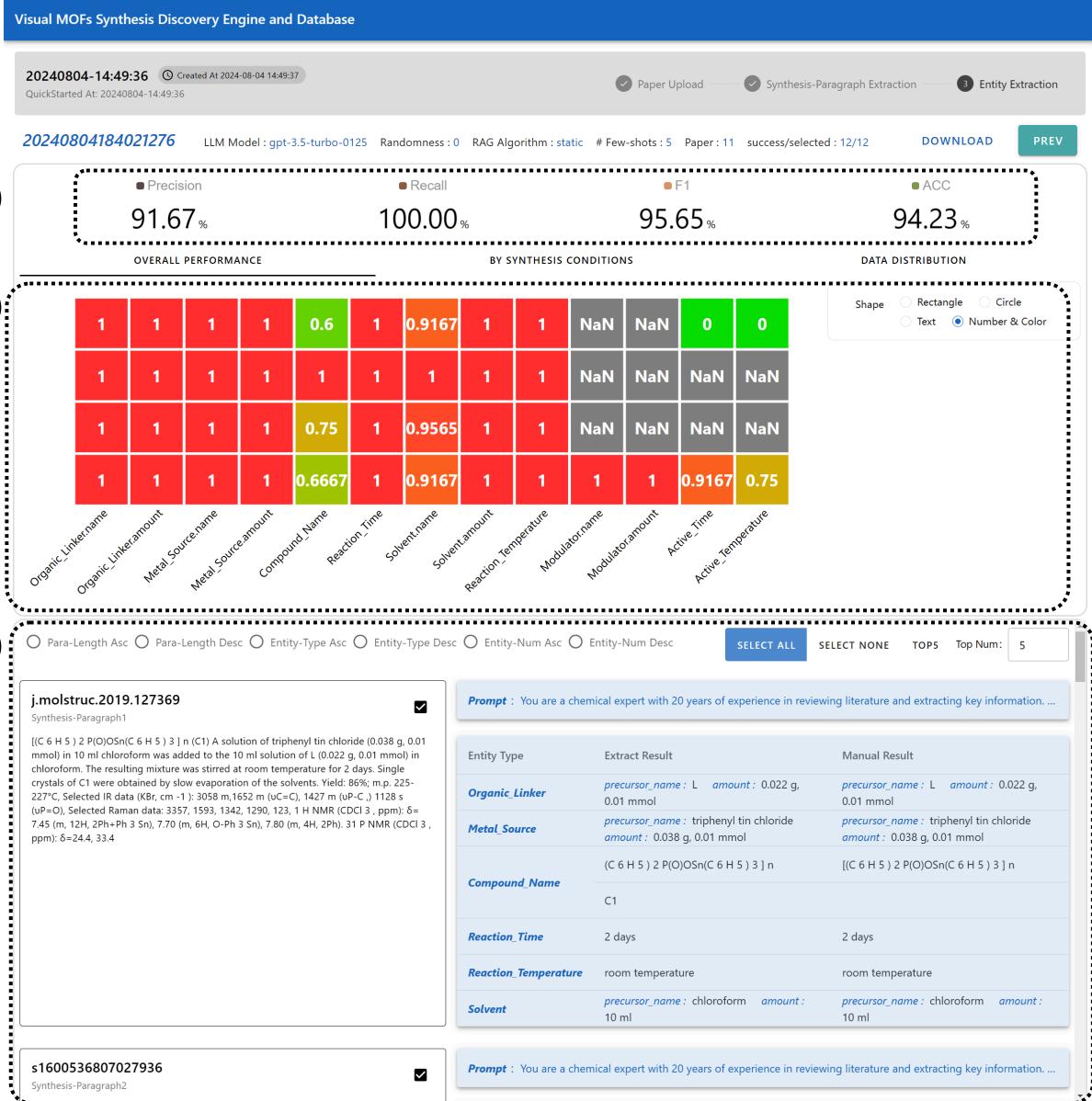


Figure S14: Visualization interface for illustrating the synthesis extraction result. (A) Overall metric panel; (B) Multi-tab detail panel (with the “OVERALL PERFORMANCE” tab selected); (C) Extraction result panel.

Visual MOFs Synthesis Discovery Engine and Database

20240804-14:49:36 (Created At 2024-08-04 14:49:37)

QuickStarted At: 20240804-14:49:36

(Paper Upload

(Synthesis-Paragraph Extraction

(Entity Extraction

20240804184021276

LLM Model : gpt-3.5-turbo-0125

Randomness : 0

RAG Algorithm : static

Few-sh...

[DOWNLOAD](#)

[PREV](#)

Precision

91.67 %

Recall

100.00 %

F1

95.65 %

ACC

94.23 %

OVERALL PERFORMANCE

Reaction Time

BY SYNTHESIS CONDITIONS

DATA DISTRIBUTION

Solvent.name

Precision: 0.917

Recall: 1.000

F1: 0.957

ACC: 0.917

- Para-Length Asc
- Para-Length Desc
- Entity-Type Asc
- Entity-Type Desc
- Entity-Num Asc
- Entity-Num Desc

[SELECT ALL](#)

[SELECT NONE](#)

Top
Num:

5

Figure S15: Synthesis extraction performance panel on each uploaded MOFs literature.

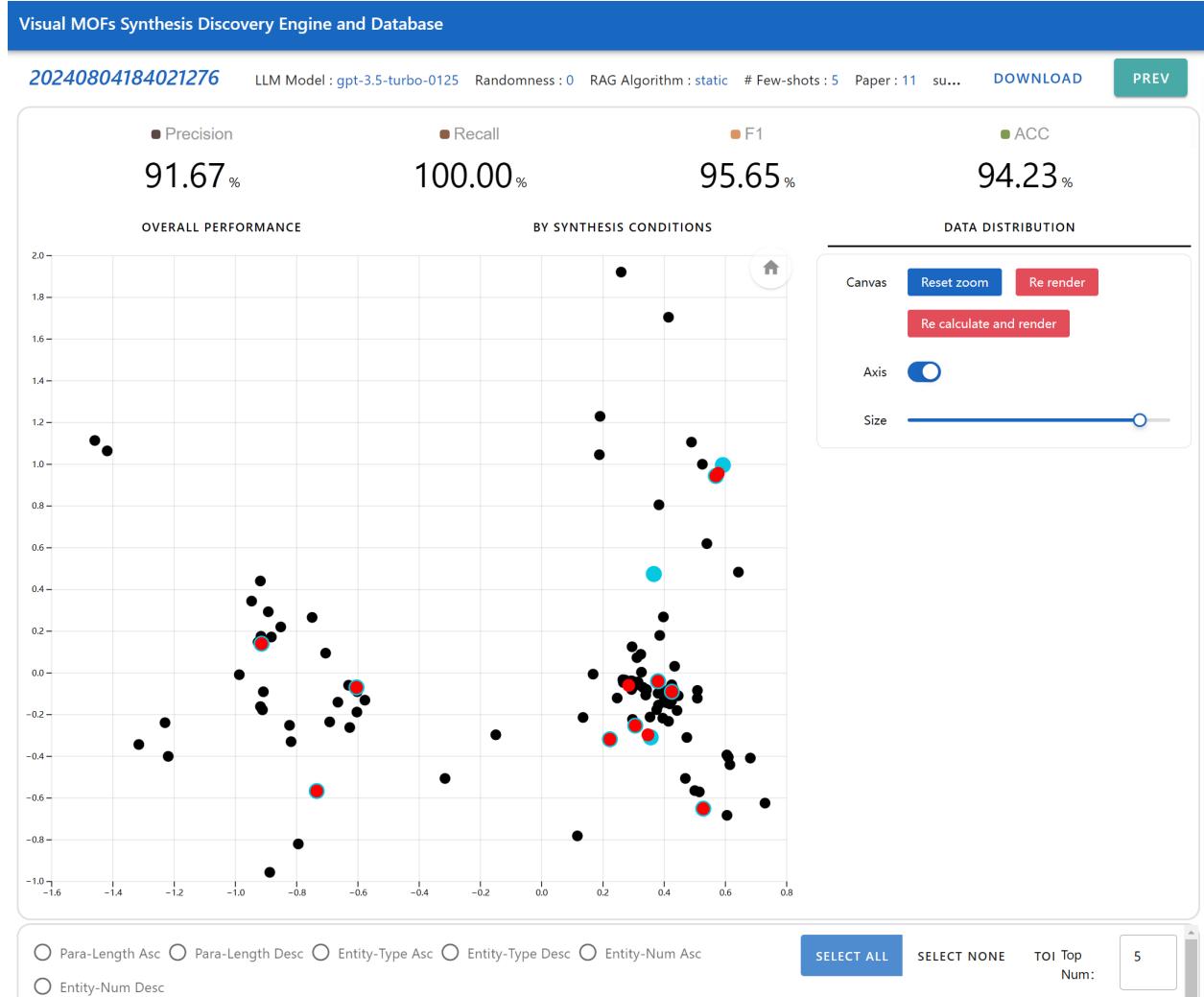


Figure S16: Visualization of the synthesis condition distribution in the 2D projection view.

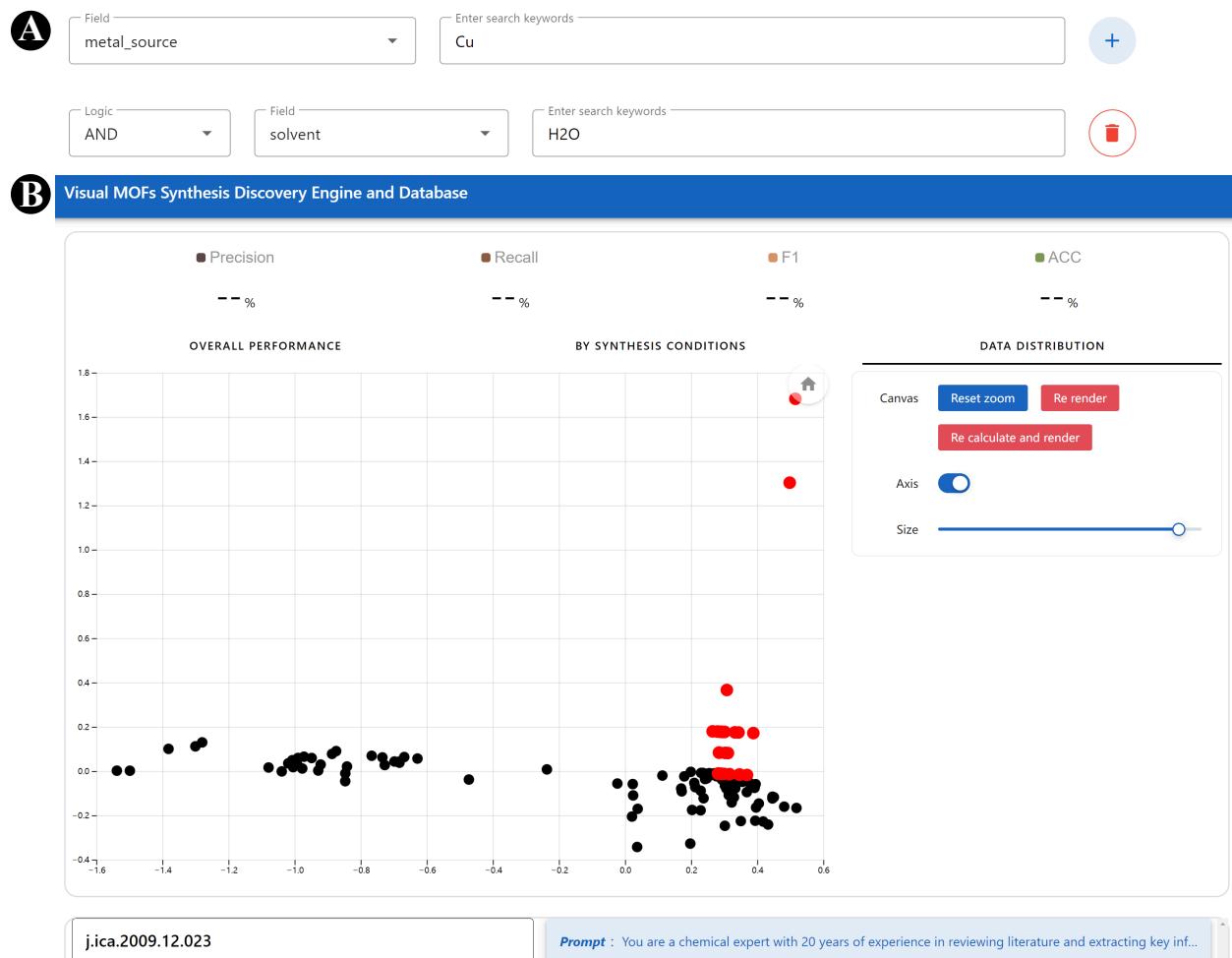


Figure S17: Visualization interface for showing the result of database retrieval. (A) Database queries; (B) Visualization of the retrieval result.

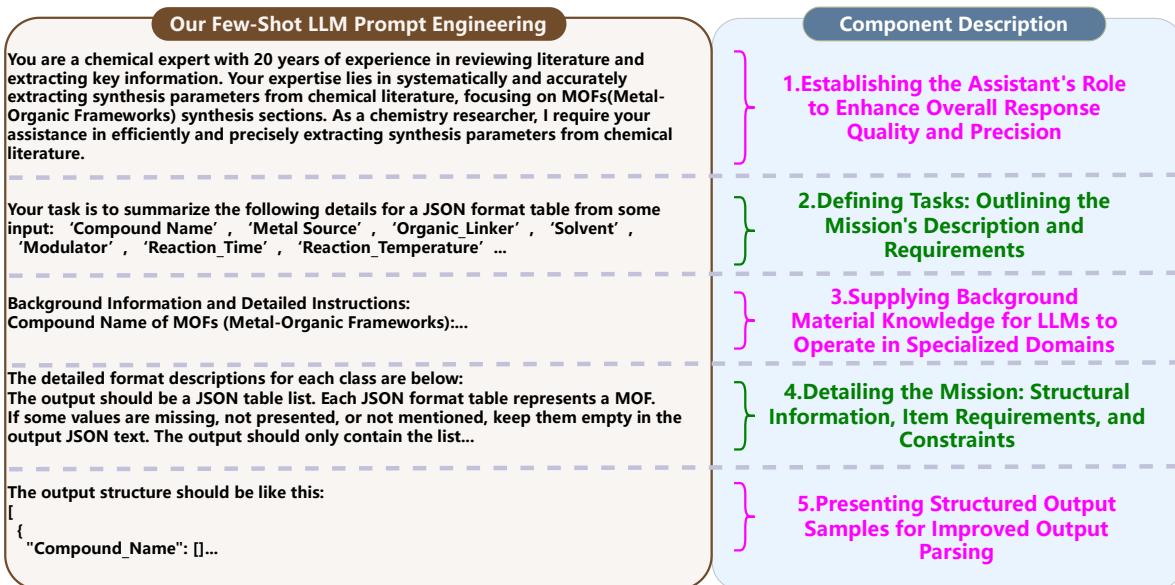


Figure S18: Structure of the LLM prompt used throughout this work.

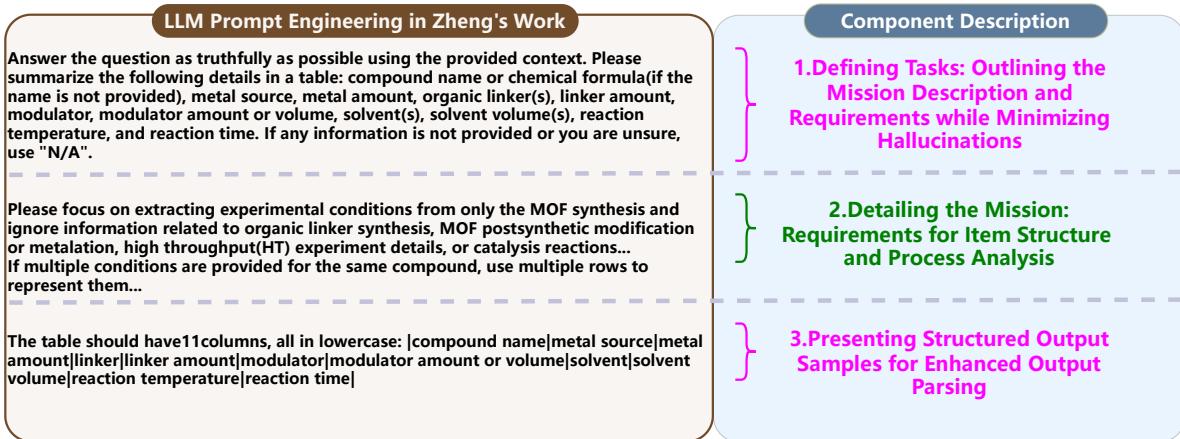


Figure S19: Structure of the LLM prompt used in Zheng's work (8).

Caption for Movie S1. Demonstration of the online LLM-based MOFs synthesis extraction engine, database retrieval, and visualization. The video shows a typical usage example composed of literature upload, synthesis paragraph detection, LLM-based synthesis route extraction, visualization-assisted extraction result analysis, as well as the information retrieval on all the extracted MOFs synthesis routes.

Caption for Data S1. Extracted MOFs synthesis route data using LLM, as well as the original human-AI annotation data and the raw literature file. Additional tabs also contain training/test data for MOFs structure inference and design. Both the data file and a full data description file are provided in the package.