RESHAPING MOFS TEXT MINING WITH A DYNAMIC MULTI-AGENTS FRAMEWORK OF LARGE LANGUAGE MODEL

Zuhong Lin[1,2,3,†], Daoyuan Ren[4,†], Kai Ran[1,2], Jing Sun[5], Songlin Yu[6], Xuefeng Bai[7], Xiaotiang Huang[1,2], Haiyang He[8], Pengxu Pan[1], Xiaohang Zhang[1], Ying Fang[1,9], Tianying Wang[1,2], Minli Wu[10], Zhanglin Li[10], Xiaochuan Zhang[11], Haipu Li[1,2,*], Jingjing Yao[1,2,*]

[1]Center for Environment and Water Resources, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, PR China

[2]Key Laboratory of Hunan Province for Water Environment and Agriculture Product Safety, Changsha, 410083, PR China

[3]School of Resources and Environment, Hunan University of Technology and Business, Changsha 410205, P.R. China

[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, PR China

[5]Faculty of Data Science, City University of Macau, 999078, Macao SAR, PR China

[6]State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China

[7]Beijing Key Laboratory for Green Catalysis and Separation, Department of Chemical Engineering, College of Materials Science and Engineering, Beijing University of Technology, Beijing 100124, China

[8]State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, 100084, PR China

[9]School of Chemical Engineering and Materials Science, Yueyang University, Yueyang 414000, PR China

[10]School of Computer Science and Engineering, Central South University, Changsha, 410083, PR China

[11]School of Software Engineering, Sun Yat-sen University, Zhuhai 519000, PR China

ABSTRACT

Accurately identifying synthesis conditions for metal–organic frameworks (MOFs) remains a critical bottleneck in materials research, as translating literature-derived knowledge into actionable insights is hindered by the unstructured and heterogeneous nature of scientific texts. Here we present MOFh6, a large language model (LLM)-based multi-agent system designed to extract, structure, and apply synthesis knowledge from diverse input formats, including raw literature and crystal codes. Built on gpt-4o-mini and fine-tuned with up to few-shot expert-annotated, MOFh6 achieves 99% accuracy in synthesis data parsing and resolves 94.1% of complex co-reference abbreviations. It processes a single full-text document in 9.6 seconds and localizes structured synthesis descriptions within 36 seconds,

[†]Zuhong Lin (lzhzzzzwill@hutb.edu.cn) and Daoyuan Ren (rendaoyuan23@mails.ucas.ac.cn) contribute equally in this work.
[*]Haipu Li (lihaipu@csu.edu.cn) and Jingjing Yao (yaojj0412@163.com) are corresponding authors.

with the cost per 100 papers reduced to $4.24, a 76% saving over existing systems. By addressing long-standing limitations in cross-paragraph semantic fusion and terminology standardization, MOFh6 reshapes the LLM-based paradigm for MOFs synthesis research, transforming static retrieval into an integrated and dynamic knowledge acquisition process. This shift bridges the gap between scientific literature and actionable synthesis design, providing a scalable framework for accelerating materials discovery.

***K*eywords** Large language model agent · Metal-organic frameworks · Data-mining · human-computer interaction · Dynamic framework

1 Introduction

Metal–organic frameworks (MOFs) are crystalline porous solids assembled from inorganic secondary building units interconnected by organic linkers into predictable, modular architectures[1, 2, 3]. Although MOFs offer great compositional and topological diversity due to their structural tunability, conventional trial-and-error synthesis remains inefficient for exploring such an expansive chemical space[4, 5, 6, 7, 8]. By 2022, the Cambridge Crystallographic Data Center (CCDC) had documented more than 110 thousand MOF structures, reflecting exponential growth and revealing the limitations of empirical discovery methods[9], underscoring the imperative shift toward data-driven intelligent material discovery paradigms.

Scientific literature underpins MOF research but is largely unstructured[10, 11, 12, 13, 14, 15, 16]. Diverse linguistic styles yield inconsistent descriptions of synthesis conditions. Cross-sentence references and ambiguous abbreviations break semantic continuity and hide key parameters. This ambiguity inflates information entropy during text-to-data conversion. Park et al. [17] extracted synthesis parameters from the literature through manual annotation combined with bidirectional long short-term memory networks and constructed a structured JSON database. Glasby et al. [18] developed the DigiMOF database based on the ChemDataExtractor tool [19], systematically integrating structural characteristics and synthesis process parameters for 15501 MOFs, including hydrothermal temperature and solvothermal time. Current natural language processing (NLP) tools still struggle with semantic complexity and require extensive coding. High-precision, flexible extraction of synthesis details remains an open challenge.

Large language models (LLMs) driven significant advancements in NLP paradigms, particularly within material science research [20, 21, 22]. Yaghi's team pioneered efficient extraction of MOF synthesis parameters using prompt engineering with GPT-3.5-turbo across 228 research articles [23]. Shi et al. further enhanced extraction accuracy via few-shot learning, demonstrating optimal GPT-4-turbo performance with only four training samples [24]. Recently, Kang et al. constructed a comprehensive database from over 40,000 literature sources by fine-tuning GPT-3.5-turbo and using prompt-driven GPT-4 [25]. Moreover, the ChatMOF system introduced by Kang illustrates the extension of LLM capabilities beyond pure text mining, integrating MOFTransformer [28] and the genetic algorithm pormake [29] for natural language-driven inverse design and performance prediction of MOFs [26, 27]. However, existing LLM-based approaches still face notable limitations, particularly in accurate cross-paragraph semantic fusion and in the standardization of highly variable chemical terminologies, underscoring the urgent need for more integrated and robust solutions.

Building upon recent advances in LLM-based MOF information mining and interactive operations, we have developed MOFh6, an enterprise-level intelligent interaction system that provides end-to-end services, from text mining to application-oriented outputs. Guided by a hierarchical architecture concept, MOFh6 was designed around strategic decision-making, technical implementation, and task execution (Figure 1), A full-chain service system was constructed with four core roles operating collaboratively. The CEO is responsible for central coordination, leveraging the Langgraph framework; the CMO oversees data management and user interaction; the COO, comprising human experts, focuses on data annotation and workflow design; the CTO, constituted by LLM agents, leads the integration and deployment of core technological capabilities. Built on this organizational synergy, MOFh6 orchestrates multi-agent collaboration to support key functionalities across its core modules. The B-end application has achieved full-text precise localization

and structured mining of specified MOF synthesis conditions, overcoming the limitations of traditional NLP in terms of insufficient semantic reference parsing ability, difficulty in unifying terminology diversity, and lack of cross-paragraph semantic fusion mechanism. This module also supports users to implement literature crawling, MOF pore structure parameter query, and crystal structure visualization functions without the need for programming knowledge. The C-end interaction module enables users to conveniently access various functions of B-end applications through natural language interfaces. Ultimately, MOFh6 addresses the core challenge of translating literature-derived synthesis knowledge into actionable insights for materials design and evaluation.

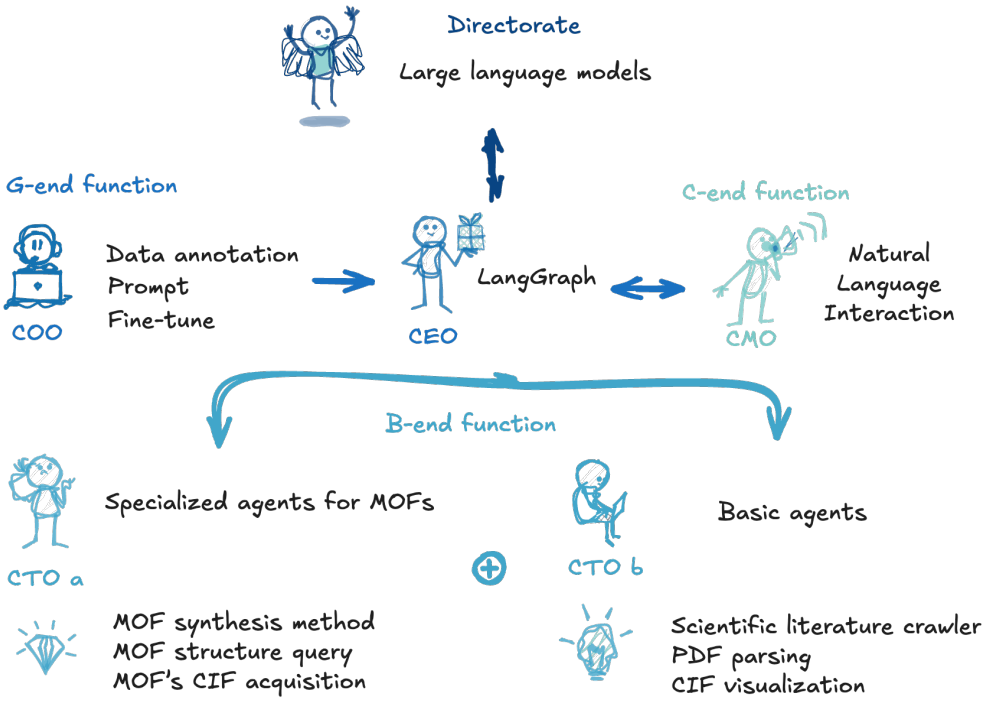

Figure 1: Enterprise architecture of MOFh6.

2 Results

2.1 Multi-tasking empowered MOFh6 pipeline

MOFh6 reconstructs the MOF knowledge extraction paradigm by unifying LLM-based semantic parsing, rule-guided refinement, and interactive crystallographic services into a coherent multi-agent workflow (Figure 2). In the synthesis process analysis Task I, the system matches the publisher of the queried MOF through a DOI routing module, dispatches institutionally authorized customized crawler scripts, and achieves compliant acquisition of target literature and standardized conversion from PDF to TXT format. The synthesis information processing unit uses fine-tuned GPT-4o-mini to build a multi-level information extraction pipeline. Firstly, MOFh6 reconstructing complete semantic contexts through the synthesis data parsing agent to obtain all descriptions about synthesis from the TXT text. Since most scientific papers provide crystallographic data for MOFs in tabular form when reporting them [30, 31], the table data parsing agent, crystal data comparison agent, and post-processor then refine and filter out the synthesis information corresponding to the queried MOF. For H_xL_x, L_xH_x, L_x ($x \geq 0$) type abbreviations referring to organic linkers, inspired by Shi et al. [24], the chemical abbreviation resolution agent uses a dual-verification mechanism of regular expressions & LLM to resolve co-references from abbreviations to full names. The final standardized synthesis parameter table covers 14 key indicators [23, 24, 32] including metal precursors, organic linkers, solvent systems, and synthesis environments

(Table S2), accommodating the processing needs of literature with different writing styles. MOFh6 simultaneously supports user-defined PDF uploads, executing an information extraction process with the same precision.

In the structure property analysis Task II, MOFh6 has deployed a metadata retrieval tool that supports users in retrieving information such as CCDC codes, chemical naming systems, and crystallographic parameters of specified MOFs through natural language queries(Figure 2). MOFh6's context-aware session management mechanism automatically resolves implicit references such as "this material" and establishes a paging index mechanism (e.g., "show 5 more results") to optimize the interactive experience of datasets. MOFh6's memory optimization strategy (which can cause it to forget when the context cache exceeds capacity) effectively avoids resource overload risks while ensuring smooth interaction. MOFh6 has a multi-attribute combination query engine that can handle complex retrieval conditions including extreme value screening and interval constraints, while integrating advanced analytical functions such as mean calculation and conditional statistics to form interactive data analysis capabilities (Text S4).

The crystal service Task III constructs a full-cycle management system for crystallographic information files (CIFs). When users request crystal structure files through CCDC codes, they can choose whether to trigger the three-dimensional visualization engine, generating an interactive interface displaying ball-and-stick models and unit cell frameworks, supporting rotation, scaling, and other multi-dimensional observation modes (Figure S50).

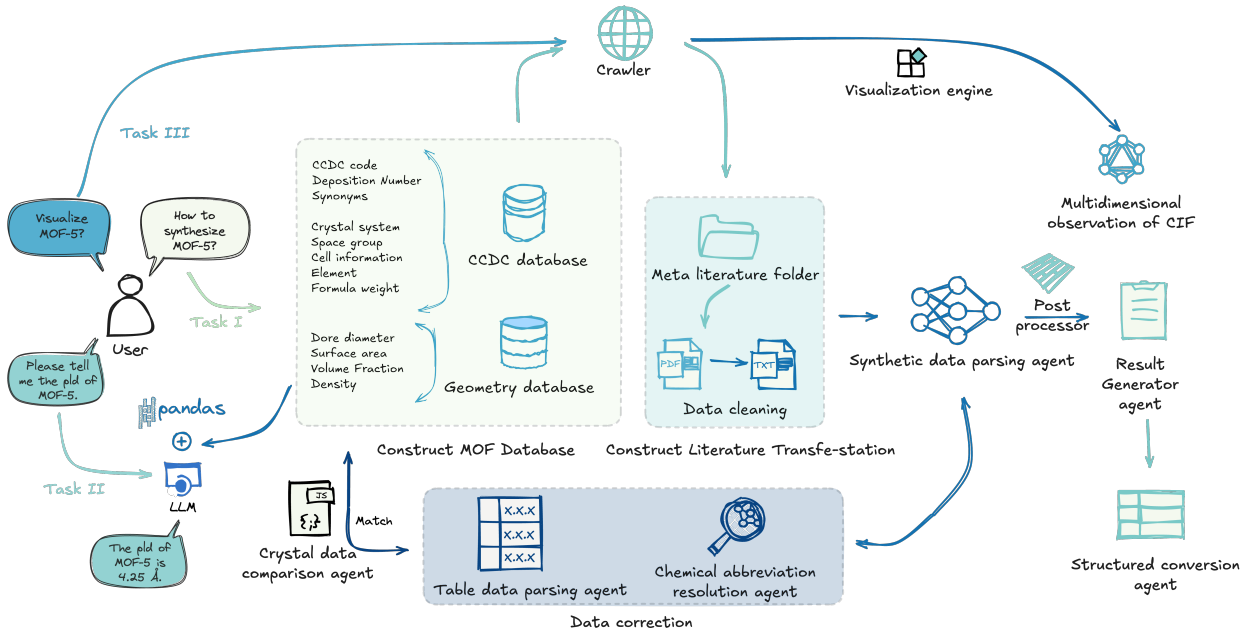

Figure 2: Overview pipeline of MOFh6. Task I operates through collaborative LLM agents; Task II integrates LLMs with a rule engine for constrained synthesis query parsing; and Task III leverages the Qt framework and the Hugging Face ecosystem to support structure visualization and CIF services.

2.2 User oriented MOFh6 Q&A Robot

Existing work shows that interactive systems based on LLMs can effectively enhance the interpretability of material data. The dialogue-based retrieval platforms constructed by Zheng et al. [23] and Kang et al. [25] reduce the cognitive threshold for non-professional users by combining structured datasets with customized LLM services. Unlike dialogue systems that rely on static databases, MOFh6 establishes a real-time, closed-loop workflow integrating data acquisition, processing, and response generation. When a user initiates a query, MOFh6 dynamically mines original literature related to the target MOFs, rather than retrieving information from pre-constructed datasets. Following data acquisition, MOFh6 uses a task-decoupling strategy to categorize both prompt engineering-based and fine-tuned LLM agents into

three coordinated stages, information extraction, logical verification, and semantic structuring. Specifically, information extraction agents consist of synthesis data parsing agents, table data processors, and chemical abbreviation resolution agents. Logical verification is handled by crystal data comparison agents and result generation agents, while semantic structuring involves post-processors and structured conversion agents. This comprehensive, three-tier quality assurance mechanism effectively reduces hallucination risks and lowers the costs associated with LLM utilization (Text S2, Figure S8-Figure S36). In doing so, MOFh6 reshapes the traditional role of LLMs from static responders to dynamically orchestrated agents for scientific knowledge extraction.

Building upon its modular agent architecture, MOFh6 further delivers interactive, low-latency query services tailored to diverse user inputs and research needs (Figure 3). Upon user login, the interactive interface automatically loads guided query templates (synthesis condition analysis, PDF processing, CIF visualization, etc.). When users submit CCDC code queries, MOFh6 retrieves the corresponding literature in TXT format through compliant crawlers. It then activates the seven-stage agent processing pipeline described in Table S1 to extract synthesis parameters. The final output consolidates key information, including the chemical name, CCDC number, commonly used abbreviations (such as ZIF-8), and detailed synthesis procedures. Additionally, MOFh6 supports one-click conversion of synthesis steps into structured Markdown tables (Figure S13-Figure S16). Since our final input is in plain text TXT format, which greatly reduces the recognition difficulty for LLMs [36, 37], the average time consumption for the entire process is 36 seconds, with an average cost of less than 0.05$ per query (Text S5).

MOFh6 adopts a dynamic context-aware architecture to implement intelligent query services, supporting users in querying MOF structural parameters in natural language (Figure S37-Figure S45). The system uses a JSON-Schema-driven semantic conversion engine that precisely maps natural language instructions to 18 structured query fields (Figure S46). The context management engine continuously tracks conversation states, records material screening sets, property constraint conditions, and historical operation trajectories, achieving semantic inheritance and logical coherence across multiple rounds of queries. MOFh6's multi-level query classification mechanism ensures the professionalism and efficiency of processing paths, with an average response time of only 3 seconds per query and an average API call cost controlled within $2.5\times10^{-4}$$. MOFh6 also provides users with convenient CIF file acquisition channels, allowing users to flexibly choose direct file downloads or use the platform's visualization functions for immediate viewing and analysis of CIF files according to research needs, thereby providing strong support for subsequent in-depth research. Additionally, MOFh6 also meets the basic multilingual interaction capabilities of LLMs, supporting users in asking questions in different natural languages (Figure S45).

2.3 Data mining statistics from data to knowledge transformation

Figure S47 presents the results of structural and bibliographic mining enabled by MOFh6 across CCDC and literature datasets, highlighting key chemical preferences, crystallographic patterns, and data source distributions that shape modern MOFs synthesis. Metal salts containing Cu, Zn, Cd, and Co elements occupy the top four usage frequencies in MOF synthesis, mainly attributed to their historical inertia, cost advantages, and stable +2 valence coordination characteristics [33] (Figure S47a). The crystallographic statistics show that low-level crystal systems such as monoclinic and triclinic occupy dominant positions, with their structural flexibility mainly stemming from the conformational diversity of carboxylic acid linkers in most MOF syntheses. Notably, these low-symmetry crystal systems often correspond to the final stable state of MOF materials after activation or oxidative treatment [34]. In contrast, the remaining medium and high-symmetry crystal systems account for only a smaller portion of the dataset, predominantly formed by rigid linkers and SBUs combinations (Figure S47b) [35]. The space group distribution shows the same differential trend as the crystal systems (Figure S47c), with MOFs having low-symmetry space groups like *P-1*, *P21/c*, and *C2/c* accounting for higher proportions than those with medium and high-symmetry space groups.

The results of literature crawling (Figure S47d) show that after 1995, most MOF structural studies were concentrated on three major publisher, the American Chemical Society (ACS), the Royal Society of Chemistry (RSC), and Elsevier.

Among them, ACS journals such as *Inorganic Chemistry* and *Crystal Growth & Design* became their main publication platforms; RSC's *CrystEngComm* and *Dalton Transactions* occupied the leading positions in their fields. Elsevier's *Inorganica Chimica Acta*, *Polyhedron*, and *Inorganic Chemistry Communication* were more popular among researchers (Figure S48). This distribution characteristic is closely related to the traditional advantages and positioning of various journals in the field of coordination chemistry, reflecting the continuous preference of MOF research groups for professional chemistry journals.

The mining results of MOF structures (Figure S49) show that the dataset exhibits a significant concentration trend in pore characteristics, with LCD mainly distributed in the 2-4 Å range and PLD concentrated in the 0-2 Å range. In the specific surface area parameters, both VSA and GSA are primarily distributed in the 0-200 m^2/cm^3 & m^2/g interval. These statistical characteristics cover most known MOF structure types, providing benchmark references for structural adaptation in different application scenarios. Through the integration of retrieval and statistical dual modes, the MOFh6 system allows users to both precisely obtain structural parameters of individual MOFs and retrieve overall distribution characteristics of the dataset in real-time, achieving fast-response structural feature analysis services.

2.4 Agents' performance of human-computer interaction verification

Numerous works have shown that LLM sample pool and model performance follow a positive correlation for performance optimization [38, 39, 40]. Figure 4a shows the impact of increasing fine-tuning sample pool on model performance in the synthesis data parsing agent. When using a pool of 50 expert-annotated samples for fine-tuning, the model achieved a prediction accuracy of 0.94 on the test set. However, considering the significant differences in writing styles among global researchers, especially the bridging reference phenomena common in scientific papers, a smaller scale training sample pool is still challenging to precisely resolve cross-textual features. To address this, we further increased the fine-tuning sample pool to 99, at which point the model prediction accuracy significantly improved to 0.98, indicating that MOFh6's synthesis data parsing agent had developed cross-textual style adaptation capabilities. Finally, as the fine-tuning sample pool expanded to 198, the model accuracy further improved to 0.99, further confirming that MOFh6's synthesis data parsing agent could effectively handle the extraction and bridging reference resolution tasks for most synthesis paragraphs in scientific literature.

In MOF synthesis literature, the naming of organic linkers presents significant complexity, with researchers frequently using abbreviations like H_xL_x, L_xH_x, L_x for co-reference expressions, posing additional challenges for automated parsing [24]. To solve this problem, we designed a specialized chemical abbreviation resolution agent based on the GPT-4o model, with its performance shown in Figure 4b. In tests on 500 MOFs across journals from five major publishers, the agent accurately identified and resolved 214 instances of co-reference abbreviations, achieving an overall success rate of 94.1%. However, differences in writing standards between publishers also caused fluctuations in co-reference resolution effectiveness, with the Wiley database performing best at a 96% success rate, while the ACS database had a reduced success rate of 81.5% due to the impact of textual expression peculiarities. Nevertheless, the resolution success rate for all databases exceeded 80%, demonstrating stable parsing capabilities especially when processing texts involving complex chemical nomenclature, confirming the agent's practical value in actual scientific research applications.

Furthermore, researchers typically use bold symbols or numbers to bridge references to numerous MOF structures in the same literature [41, 42, 43, 44]. While these abstract expressions are concise, they may affect the quick understanding of newcomers to the field. Our synthesis data parsing agent can effectively resolve bridging reference phenomena, achieving complete semantic restoration. As shown in Figure 4c, when users submit queries with specific CCDC codes (such as ABAYUY, which includes three crystals in the literature [45]), the system first automatically extracts multiple crystal synthesis paragraphs from the original literature, then generates complete semantic descriptions through bridging reference resolution. These structured data are subsequently processed further by the table data parsing agent and crystal data comparison agent, ultimately achieving precise localization of the synthesis description corresponding to the target

crystal code (Figure S52). System evaluations show that as the sample pool expanded from 25 to 100, MOFh6 agents' comprehensive accuracy in parsing synthesis information from literature across five major publishers fluctuated slightly between 0.94 and 0.93, maintaining overall performance stability within $\pm 1\%$. Specifically, the Elsevier database exhibited extremely strong robustness, with cosine similarity between post-text parsing and manually annotated text remaining stable between 0.97 and 0.98, while the RSC database performed relatively weaker, maintaining cosine similarity between 0.89 and 0.90. The parsing accuracy for other publishers generally remained above 0.91. Overall, MOFh6 possesses cross-scale sample adaptation capabilities and can effectively achieve accurate restoration and localization of synthesis paragraphs corresponding to given CCDC codes.

To enhance the readability and practical utility of information on MOF synthesis conditions, we structured the chemical composition (Chemicals), reaction conditions (Conditions), and crystal characteristics (Crystallization) using the 3C format (Figure 4d). Regarding specific chemical components, the identification of metal salts exhibited exceptional performance, largely due to the clear correspondence between chemical formulas and their full names, typically provided explicitly in literature sources. This clarity resulted in accuracy, precision, recall, and F1 scores all exceeding 0.99. In contrast, organic linkers presented higher semantic complexity, particularly between structurally similar isomers, such as "(R)-2,2'-dihydroxy-1,1'-binaphthyl-4,4',6,6'-tetrabenzoate" and "(R)-2,2'-diethoxy-1,1'-binaphthyl-4,4',6,6'-tetrabenzoate" [46]. This complexity slightly impacted identification accuracy and precision, lowering these scores to approximately 0.94, although recall and F1 scores remained above 0.94. Additionally, additives used for pH adjustment (e.g., acids, bases, and triethylamine) [23, 47] and natural language descriptions of solvents are relatively standardized, resulting in stable identification performance with comprehensive scores exceeding 0.93. While the overall identification performance for quantities of metal salts, organic linkers, and additives remained excellent, with scores consistently above 0.90, challenges arose when processing combinations involving multiple solvents or repeated mentions [48]. Due to inherent limitations in arithmetic reasoning capabilities of LLMs, slight deviations occurred in the calculation of quantities, marginally reducing the comprehensive score to approximately 0.89. In terms of Conditions, the structured conversion agent demonstrated stable and precise parsing performance for single-parameter conditions (e.g., pH, synthesis temperature, synthesis time), achieving comprehensive scores above 0.92. However, the accuracy was somewhat diminished when dealing with multi-stage equipment descriptions, such as "initial hydrothermal synthesis in a Teflon reactor followed by evaporation crystallization in a beaker" [47], resulting in a slightly reduced comprehensive performance score of approximately 0.83. Regarding Crystallization, the structured conversion agent effectively extracted crystal morphology descriptions, achieving a comprehensive accuracy above 0.85. Its capability in recognizing yield, which is quantitative information, was particularly notable, reaching precision levels above 0.91. This performance underscores the agent's strong ability to handle information containing both qualitative and quantitative elements. Overall, the MOFh6 structured conversion agent demonstrates commendable stability and adaptability when processing complex and diverse chemical texts.

3 Conclusion

In this work, we developed MOFh6, an end-to-end, enterprise-level intelligent system that enables full-chain interaction for MOFs synthesis knowledge extraction and application. By orchestrating multiple agents in a collaborative framework, MOFh6 reshapes the conventional pipeline of MOF informatics, integrating dynamic literature acquisition, context-aware parsing, and structured knowledge generation into a unified workflow. This design significantly lowers the cognitive barrier for non-specialist users and reduces the data mining costs. Experimental results show that a fine-tuning strategy based on 198 sets of expert annotations achieved a synthesis parameter extraction accuracy of 0.99, effectively overcoming the challenge of cross-literature bridging reference resolution. For the heterogeneous co-reference phenomenon of ligand abbreviations in chemical texts, the system demonstrated an average resolution success rate of 94.1% in tests across literature from five major publishers. The system maintained a precision of 0.93±0.01 for synthesis descriptions of specified MOFs when the sample scale was expanded. Based on 3C structuring format, the system achieved accuracy, recall, and F1 scores above 0.8 for various structured descriptions of MOF synthesis,

establishing a standardized data pipeline for high-throughput material development. The successful validation of MOFh6 reshapes the feasibility of the LLM agents and rule engine-driven paradigm for acquiring synthesis methods of MOFs and broader material systems, with its modular architecture enabling an intelligent closed loop from database mining to synthesis protocol generation. However, the current system is still limited by third-party tool compatibility in terms of dynamic data source integration. In the future, it will achieve seamless integration of CSD-Python-API and the LLM ecosystem through architectural upgrades, promoting the evolution of material development towards a fully digitalized stage of flexible human-machine interaction.

4 Methods

4.1 Construction of crystallographic dataset

The MOF crystal database constructed in this study ultimately formed a standardized dataset containing 43766 crystal structures. Through the CSD-Python-API interface of the Cambridge Crystallographic Data Centre (CCDC) [49], we achieved automated extraction and structured processing of crystal information, generating JSON format metadata including multiple dimensional features such as MOFs' CCDC codes, numbers, chemical names, abbreviations, DOI numbers and corresponding URLs of publications, as well as space group information, crystal information, unit cell lengths, angles, elemental composition, and molecular weight.

Simultaneously, through the open-source software Zeo++ [50], using N_2 with a kinetic radius of 1.82 Å as a molecular probe, we calculated core physical parameters of MOFs including pore limiting diameter, largest cavity diameter, crystal density, volumetric accessible surface area, gravimetric accessible surface area, and volume fraction, constructing a physical parameter database for MOFs.

4.2 Literature data mining

We strictly adhered to academic publishing ethical guidelines and digital copyright management standards in our literature acquisition work. Relying on institutional subscriptions to mainstream publishing platforms such as Elsevier, Wiley, American Chemical Society (ACS), Royal Society of Chemistry (RSC), and Springer, we systematically constructed a target literature dataset within the range of institutionally authorized IPs using an automated literature mining system. Compared to existing research limited by platform interfaces or unable to fully obtain supporting materials [25, 51], MOFh6's crawler module achieves deep parsing of full-text formats, specifically addressing the disciplinary characteristic where synthesis descriptions are often distributed in supporting materials. By completely acquiring the main body of the literature and supplementary material content, it constructs a complete underlying data architecture for fine-grained semantic analysis by LLMs, effectively solving the parsing challenge of fragmented distribution of key information in scientific literature.

4.3 Literature processing system framework

The core of our text processing system is based on LangGraph* constructing a three-layer processing architecture in the form of a directed state graph, achieving collaborative processing of multiple modules through declarative workflows (Table S1). The first layer is the data collection and preprocessing structure, where the synthesis data parsing agent parses input documents through fine-tuned LLMs, focusing on solving bridging reference problems in synthesis paragraphs (e.g., context-dependent expressions like "this method," "above conditions"), generating synthesis description texts with complete semantics. The table data parsing agent, running in parallel, adopts LLM prompt engineering to precisely identify crystallographic parameter tables in documents (including unit cell parameters, space groups, etc.) and convert them into standardized JSON format. The second layer is the core data processing structure, where the crystal data comparison agent establishes associative mapping between preprocessed data, matching crystallographic parameters in

*https://github.com/langchain-ai/langgraph

tables with specific crystals described in synthesis texts through feature matching algorithms. For chemical abbreviation patterns like H_xL_x, L_xH_x, L_x ($x \geq 0$) in the literature, the chemical abbreviation resolution agent adopts a dual verification mechanism combining regular expression rule libraries with constrained prompt engineering, achieving normalized naming of chemical entities while controlling LLM hallucination risks. The third layer is the structured output structure, where the result generation agent constructs a document-level co-reference resolution pipeline based on the BM25 algorithm, integrating previous processing results to generate standardized MOF synthesis description texts. The post-processor executes multi-file output management, generating independent sub-files for each input document. The structured conversion agent converts the final text into a Markdown table containing structured fields such as synthesis conditions and crystal parameters through predefined templates (Text S1).

The system achieves state transfer between nodes through directed edges, with each processing module supporting independent configuration and extension, demonstrating stable processing performance and effectively solving key problems in chemical literature processing such as reference resolution and cross-modal data association. This architecture provides a verifiable technical path for automated processing of textual data in the field of materials science.

4.4 Query & Answer system framework

To help non-professionals quickly obtain structural and synthesis information about MOFs, we developed an intelligent question-answering system based on natural language interaction. This system adopts a three-layer architectural design, achieving efficient information retrieval through intelligent data parsing and visualization technology, with conversational query and interaction capabilities.

At the data architecture level, the system constructs a standardized preprocessing workflow, implementing efficient positioning of raw data through the Pandas data framework, and adopting a dual strategy combining LLMs with dynamic field mapping. This design effectively avoids the hallucination problems that might be produced by a single LLM, while supporting users in querying using any structural description terminology. For complex query scenarios, the system developed a context-aware query parsing mechanism, achieving semantic association understanding in conversational scenarios through multi-level mapping dictionaries and persistent context management. The core parsing module adopts a primary-backup dual-engine design, with the main parser completing semantic conversion from natural language to JSON structures through LLMs, and the backup parser ensuring the reliability of key parameter extraction through function calls, accurately identifying seven types of query modes including direct queries, range queries, and comparison queries.

The interaction system adopts a state-management command-line interface, integrating diversified functions such as PDF parsing, CIF downloading, and visualization. The PDF processing module implements parallel extraction of document structure and content based on the PyMuPDF engine, and can effectively process special symbols and layout anomalies in academic literature through a text cleaning process optimized by regular expressions. For literature materials uploaded by users, the system preserves paragraph semantic structures through standardized processing, laying the foundation for subsequent extraction of synthesis information.

For specific MOF synthesis information queries, the system deploys a crawler router. When users submit CCDC numbers, the DOI routing module automatically identifies the corresponding publisher's DOI and URL, triggering customized crawlers to obtain full text. The crystal structure visualization module adopts a dual-engine architecture of pymatgen and py3Dmol, achieving precise conversion from CIF files to interactive three-dimensional models. The visualization interface, developed based on the PyQt5 framework, clearly displays the channel topology and coordination modes of MOF materials through a combination of ball-and-stick models and unit cell parameters, supporting multi-dimensional observation including rotation and scaling.

4.5 Prompt engineering and fine-tuning of model

In LLM applications, prompt engineering has been proven to be a key technical strategy. By setting clear instructions, constructing contextual constraint conditions, and providing a small number of sample cases, it can effectively suppress model hallucination phenomena and significantly improve information extraction precision and data generation quality [23, 24, 52, 53]. Following LangGPT rules [54], we constructed high-quality prompts with clear role instructions and few-shot examples for different modules to maximize model performance (Text S2).

Fine-tuning can significantly reduce model operating costs and increase model performance [32, 55]. In the model fine-tuning stage, we maximized model performance by optimizing training sample selection strategies and ensuring sufficient training scale (token count). Technically, we integrated GPT-4o-mini and GPT-4o dual-model architecture through API interfaces, with GPT-4o-mini having expanded supervised fine-tuning (SFT) functionality, enabling scenario adaptability optimization for private datasets. In the specific implementation process, experimental data was normalized into JSONL format and uploaded to the OpenAI cloud service platform, then initiating customized fine-tuning tasks. After completing fine-tuning training, the model could be deployed to actual application scenarios. To ensure model training effectiveness, we constructed datasets containing 50, 99, and 198 sets of expert-annotated texts, randomly allocating them to training, validation, and test sets in an 8:1:1 ratio.

4.6 Index evaluation

We adopted a human-machine interaction mixed evaluation strategy to verify system performance. For human evaluation, we formed a group of four human experts at the master's level or above, randomly selecting 100 MOFs with corresponding literature from each of the five major publishers for annotation. Group members ensured annotation reliability through a cross-annotation protocol. At the algorithmic evaluation level, considering the special nature of chemical semantics, we adopted the PubMedBERT model [56] for evaluation. The PubMedBERT model itself is a BERT model trained from scratch on a large amount of biomedical literature, excelling at capturing semantic relationships between chemical professional terminology and complex chemical descriptions. For the structured parameter extraction task of MOF synthesis, a dual-model collaborative encoding strategy was adopted. We used the PubMedBERT model for professional naming fields such as chemical names, and the all-mpnet-base-v2 model for general structured fields for semantic representation (Text S3, Table S2) (eq. 1), calculating cosine similarity through batch processing of size 32 and mixed precision. Based on the complexity of chemical expressions, the threshold was set at 0.90.

$$\cos(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Where A and B represent different embedding vectors, and n is the dimension of the vectors. Further we calculated accuracy, precision, recall, and F1 score (eq. 2-eq. 5).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Where TP (True Positive) indicates that the human expert annotation (gold standard) and LLM judgment on the target parameters of MOF synthesis achieve complete semantic consistency; FP (False Positive) indicates that the parameter values or descriptions output by the LLM have significant deviations from the gold standard, or erroneously identify non-synthesis-related parameters; the LLM's identification results do not match human annotations; FN (False Negative)

indicates that the LLM fails to detect key parameter entries clearly marked in the gold standard, outputting empty values; TN (True Negative) indicates that when there are indeed no relevant synthesis parameters in the target literature, neither LLM nor experts make annotations.

Author contributions

H.P.L. supervised this work. H.P.L., J.J.Y., Z.H.L., and D.Y.R conceived this idea. Z.H.L., and D.Y.R designed the entire study. Z.H.L., and D.Y.R implemented technical specifications. H.Y.H., J.S. provided technical support. T.Y.W., M.L.W., Z.L.L., X.C.Z. collected the meta data. S.L.Y. and X.F.B checked the front-end framework. Z.H.L., K.R., P.X.P., X.H.Z. annotated, and checked the training data. D.Y.R conducted benchmark testing on the model. Z.H.L. wrote the initial manuscript. H.P.L., J.J.Y., Y.F. reviewed and improved this manuscript. All authors contributed to the analysis of the results. All authors have read and approved the final manuscript.

Competing interests

The other authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this manuscript, the authors used AI-based language models to assist with grammar correction, style refinement, and improvement of sentence clarity. All scientific content, analysis, data interpretation, and final decisions were made solely by the authors, who take full responsibility for the integrity and accuracy of the manuscript.

Acknowledgments

The authors acknowledge financial was supported by the National Key Research and Development Program of China (2023YFC3207000). And we are grateful for resources from the High-Performance Computing Center of Central South University.

References

[1] S. Natarajan and P. Mahata. Metal-organic framework structures–how closely are they related to classical inorganic structures? *Chemical Society Reviews*, 38(8):2304–18, 2009.

[2] Hao Jiang, Dalal Alezi, and Mohamed Eddaoudi. A reticular chemistry guide for the design of periodic solids. *Nature Reviews Materials*, 6(6):466–487, 2021.

[3] O. M. Yaghi, Guangming Li, and Hailian Li. Selective binding and removal of guests in a microporous metal-organic framework. *Nature*, 378:703–706, 1995.

[4] Alec Wang, Madeline Walden, Romy Ettlinger, Fabian Kiessling, Jeremiah Gassensmith, Twan Lammers, Stefan Wuttke, and Quim Peña. Biomedical metal–organic framework materials: Perspectives and challenges. *Advanced Functional Materials*, 34, 2023.

[5] Prasad Lokhande, Sahana Kulkarni, Sandip Chakrabarti, Habib Pathan, Monika Sindhu, Deepak Kumar, Jashanpreet Singh, Anupam Kumar, Yogendra Mishra, Dana-Cristina Toncu, Mikael Syväjärvi, Ajit Sharma, and Ashutosh Tiwari. The progress and roadmap of metal–organic frameworks for high-performance supercapacitors. *Coordination Chemistry Reviews*, 473:214771, 2022.

[6] Dian Zhao, Xinxin Wang, Lianglan Yue, Yabing He, and Banglin Chen. Porous metal-organic frameworks for hydrogen storage. *Chemical Communications*, 58, 2022.

[7] Zuhong Lin, Hui Cai, Hongjia Peng, Ying Fang, Pengxu Pan, Haipu Li, Ying Yang, and Jingjing Yao. Enhancing arsenate removal through interpretable machine learning guiding the modular design of metal–organic frameworks. *Chemical Engineering Journal*, 497:155058, 2024.

[8] Zuhong Lin, Jiarong Chen, Ying Fang, Shi-hai Deng, Haipu Li, Ying Yang, and Jingjing Yao. Rapidly tailor metal–organic frameworks for arsenate removal using graph convolutional neural networks. *Separation and Purification Technology*, 354:129334, 2025.

[9] Guangtong Hai and Haihui Wang. Theoretical studies of metal-organic frameworks: Calculation methods and applications in catalysis, gas separation, and energy storage. *Coordination Chemistry Reviews*, 469, 2022.

[10] Jibin Zhou, Duiping Liu, Mao Ye, and Zhongmin Liu. Data-driven prediction of minimum fluidization velocity in gas-fluidized beds using data extracted by text mining. *Industrial & Engineering Chemistry Research*, 60(37):13727–13739, 2021.

[11] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, and A. Jain. Structured information extraction from scientific text with large language models. *Nat Commun*, 15(1):1418, 2024.

[12] Xuefeng Bai, Song He, Yi Li, Yabo Xie, Xin Zhang, Wenli Du, and Jian-Rong Li. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

[13] Junkil Park, Honghui Kim, Yeonghun Kang, Yunsung Lim, and Jihan Kim. From data to discovery: Recent trends of machine learning in metal–organic frameworks. *JACS Au*, 4(10):3727–3743, 2024.

[14] Zuhong Lin, Siyuan Pan, Haipu Li, and Jingjing Yao. Unveiling similarities and differences in oxidation processes of oxidants and derived reactive oxygen species through machine learning interpretation of oxidation rate constants. *Separation and Purification Technology*, 340:126649, 2024.

[15] Jingjing Yao, Haipu Li, and Hui Ying Yang. Predicting adsorption capacity of pharmaceuticals and personal care products on long-term aged microplastics using machine learning. *Journal of Hazardous Materials*, 458:131963, 2023.

[16] Fengyu Zhang, Xinchao Su, Aoli Tan, Jingjing Yao, and Haipu Li. Prediction of research octane number loss and sulfur content in gasoline refining using machine learning. *Energy*, 261:124823, 2022.

[17] Hyunsoo Park, Yeonghun Kang, Wonyoung Choe, and Jihan Kim. Mining insights on metal–organic framework synthesis from scientific literature texts. *Journal of Chemical Information and Modeling*, 62(5):1190–1198, 2022.

[18] L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko, S. M. Moosavi, J. L. Cordiner, J. C. Cole, and P. Z. Moghadam. Digimof: A database of metal-organic framework synthesis information generated via text mining. *Chem Mater*, 35(11):4510–4524, 2023.

[19] Matthew C. Swain and Jacqueline 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] Songlin Yu, Nian Ran, and Jianjun Liu. Large-language models: The game-changers for materials science research. *Artificial Intelligence Chemistry*, 2(2):100076, 2024.

[21] Yiming Zhao, Yongjia Zhao, Jian Wang, and Zhuo Wang. Artificial intelligence meets laboratory automation in discovery and synthesis of metal–organic frameworks: A review. *Industrial & Engineering Chemistry Research*, 64(9):4637–4668, 2025.

[22] C. W. Kosonocky, C. O. Wilke, E. M. Marcotte, and A. D. Ellington. Mining patents with large language models elucidates the chemical function landscape. *Digit Discov*, 3(6):1150–1159, 2024.

[23] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar 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.

[24] Lei Shi, Zhimeng Liu, Yi Yang, Weize Wu, Yuyang Zhang, Hongbo Zhang, Jing Lin, Siyu Wu, Zihan Chen, and Ruiming Li. Llm-based mofs synthesis condition extraction using few-shot demonstrations. *arXiv preprint arXiv:2408.04665*, 2024.

[25] Yeonghun Kang, Wonseok Lee, Taeun Bae, Seunghee Han, Huiwon Jang, and Jihan Kim. Harnessing large language models to collect and analyze metal–organic framework property data set. *Journal of the American Chemical Society*, 2025.

[26] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[27] Yeonghun Kang and Jihan Kim. Chatmof: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nature Communications*, 15(1):4705, 2024.

[28] Yeonghun Kang, Hyunsoo Park, Berend Smit, and Jihan Kim. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence*, 5(3):309–318, 2023.

[29] S. Lee, B. Kim, H. Cho, H. Lee, S. Y. Lee, E. S. Cho, and J. Kim. Computational screening of trillions of metal-organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021.

[30] Sihai Yang, Xiang Lin, William Lewis, Mikhail Suyetin, Elena Bichoutskaia, Julia E. Parker, Chiu C. Tang, David R. Allan, Pierre J. Rizkallah, Peter Hubberstey, Neil R. Champness, K. Mark Thomas, Alexander J. Blake, and Martin Schröder. A partially interpenetrated metal–organic framework for selective hysteretic sorption of carbon dioxide. *Nature Materials*, 11(8):710–716, 2012.

[31] Hiroshi Sato, Ryotaro Matsuda, Kunihisa Sugimoto, Masaki Takata, and Susumu Kitagawa. Photoactivation of a nanoporous crystal for on-demand guest trapping and conversion. *Nature Materials*, 9(8):661–666, 2010.

[32] Wei Zhang, Qinggong Wang, Xiangtai Kong, Jiacheng Xiong, Shengkun Ni, Duanhua Cao, Buying Niu, Mingan Chen, Yameng Li, Runze Zhang, Yitian Wang, Lehan Zhang, Xutong Li, Zhaoping Xiong, Qian Shi, Ziming Huang, Zunyun Fu, and Mingyue Zheng. Fine-tuning large language models for chemical text mining. *Chemical Science*, 2024.

[33] Brooke P. Carpenter, A. Rain Talosig, Ben Rose, Giuseppe Di Palma, and Joseph P. Patterson. Understanding and controlling the nucleation and growth of metal–organic frameworks. *Chemical Society Reviews*, 52(20):6918–6937, 2023.

[34] Dipendu Saha and Shuguang Deng. Structural stability of metal organic framework mof-177. *The Journal of Physical Chemistry Letters*, 1(1):73–78, 2010.

[35] Daniel J. O'Hearn, Alankriti Bajpai, and Michael J. Zaworotko. The "chemistree" of porous coordination networks: Taxonomic classification of porous solids to guide crystal engineering studies. *Small*, 17(22):2006351, 2021.

[36] Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

[37] Shican Wu, Xiao Ma, Dehui Luo, Lulu Li, Xiangcheng Shi, Xin Chang, Xiaoyun Lin, Ran Luo, Chunlei Pei, and Changying Du. Automated review generation method based on large language models. *arXiv preprint arXiv:2407.20906*, 2024.

[38] H. Liu, H. Yin, Z. Luo, and X. Wang. Integrating chemistry knowledge in large language models via prompt engineering. *Synth Syst Biotechnol*, 10(1):23–38, 2025.

[39] Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. *arXiv preprint arXiv:2402.13904*, 2024.

[40] Chuhan Wu and Ruiming Tang. Performance law of large language models. *arXiv preprint arXiv:2408.09895*, 2024.

[41] Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes, and O. M. Yaghi. A gpt-4 reticular chemist for guiding mof discovery. *Angew Chem Int Ed Engl*, 62(46):e202311983, 2023.

[42] Takehide Kawamichi, Tsuyoshi Haneda, Masaki Kawano, and Makoto Fujita. X-ray observation of a transient hemiaminal trapped in a porous network. *Nature*, 461(7264):633–635, 2009.

[43] Pei-Qin Liao, Ai-Xin Zhu, Wei-Xiong Zhang, Jie-Peng Zhang, and Xiao-Ming Chen. Self-catalysed aerobic oxidization of organic linker in porous crystal for on-demand regulation of sorption behaviours. *Nature Communications*, 6(1):6350, 2015.

[44] Yingjie Zhang, David Collison, Francis R. Livens, Madeleine Helliwell, Frank Heatley, Annie K. Powell, Sigrid Wocadlo, and Harry Eccles. Synthesis and characterisation of uranyl substituted malonato complexes: Part ii: 13c cpmas nmr spectroscopy related to structural diversity. *Polyhedron*, 21(1):81–96, 2002.

[45] Xiaoju Li, Rong Cao, Daofeng Sun, Wenhua Bi, Yanqin Wang, Xing Li, and Maochun Hong. Syntheses and characterizations of zinc(ii) compounds containing three-dimensional interpenetrating diamondoid networks constructed by mixed ligands. *Crystal Growth & Design*, 4(4):775–780, 2004.

[46] Liqing Ma, Chuan-De Wu, Marcela M. Wanderley, and Wenbin Lin. Single-crystal to single-crystal cross-linking of an interpenetrating chiral metal–organic framework and implications in asymmetric catalysis. *Angewandte Chemie International Edition*, 49(44):8244–8248, 2010.

[47] Jian-Bin Lin, Jie-Peng Zhang, and Xiao-Ming Chen. Nonclassical active site for enhanced gas sorption in porous coordination polymer. *Journal of the American Chemical Society*, 132(19):6654–6656, 2010.

[48] Young Eun Cheon and Myunghyun Paik Suh. Multifunctional fourfold interpenetrating diamondoid network: Gas separation and fabrication of palladium nanoparticles. *Chemistry – A European Journal*, 14(13):3961–3967, 2008.

[49] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward. The cambridge structural database. *Acta Crystallogr B Struct Sci Cryst Eng Mater*, 72(Pt 2):171–9, 2016.

[50] Thomas F. Willems, Chris H. Rycroft, Michaeel Kazi, Juan C. Meza, and Maciej Haranczyk. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*, 149(1):134–141, 2012.

[51] N. Smith, X. Yuan, C. Melissinos, and G. Moghe. Funcfetch: An llm-assisted workflow enables mining thousands of enzyme-substrate interactions from published manuscripts. *Bioinformatics*, 2024.

[52] X. Bai, Y. Xie, X. Zhang, H. Han, and J. R. Li. Evaluation of open-source large language models for metal-organic frameworks research. *Journal of Chemical Information and Modeling*, 64(13):4958–4965, 2024.

[53] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[54] Ming Wang, Yuanzhong Liu, Xiaoyu Liang, Songlian Li, Yijie Huang, Xiaoming Zhang, Sijia Shen, Chaofeng Guan, Daling Wang, and Shi Feng. Langgpt: Rethinking structured reusable prompt design framework for llms from the programming language. *arXiv preprint arXiv:2402.16929*, 2024.

[55] Z. Zheng, A. H. Alawadhi, S. Chheda, S. E. Neumann, N. Rampal, S. Liu, H. L. Nguyen, Y. H. Lin, Z. Rong, J. I. Siepmann, L. Gagliardi, A. Anandkumar, C. Borgs, J. T. Chayes, and O. M. Yaghi. Shaping the water-harvesting behavior of metal-organic frameworks aided by fine-tuned gpt models. *J Am Chem Soc*, 145(51):28284–28295, 2023.

[56] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3:1–23, 2022.

[57] C. M. Castro Nascimento and A. S. Pimentel. Do large language models understand chemistry? a conversation with chatgpt. *J Chem Inf Model*, 63(6):1649–1655, 2023.

[58] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024.

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[60] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*, 2024.

[61] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020.

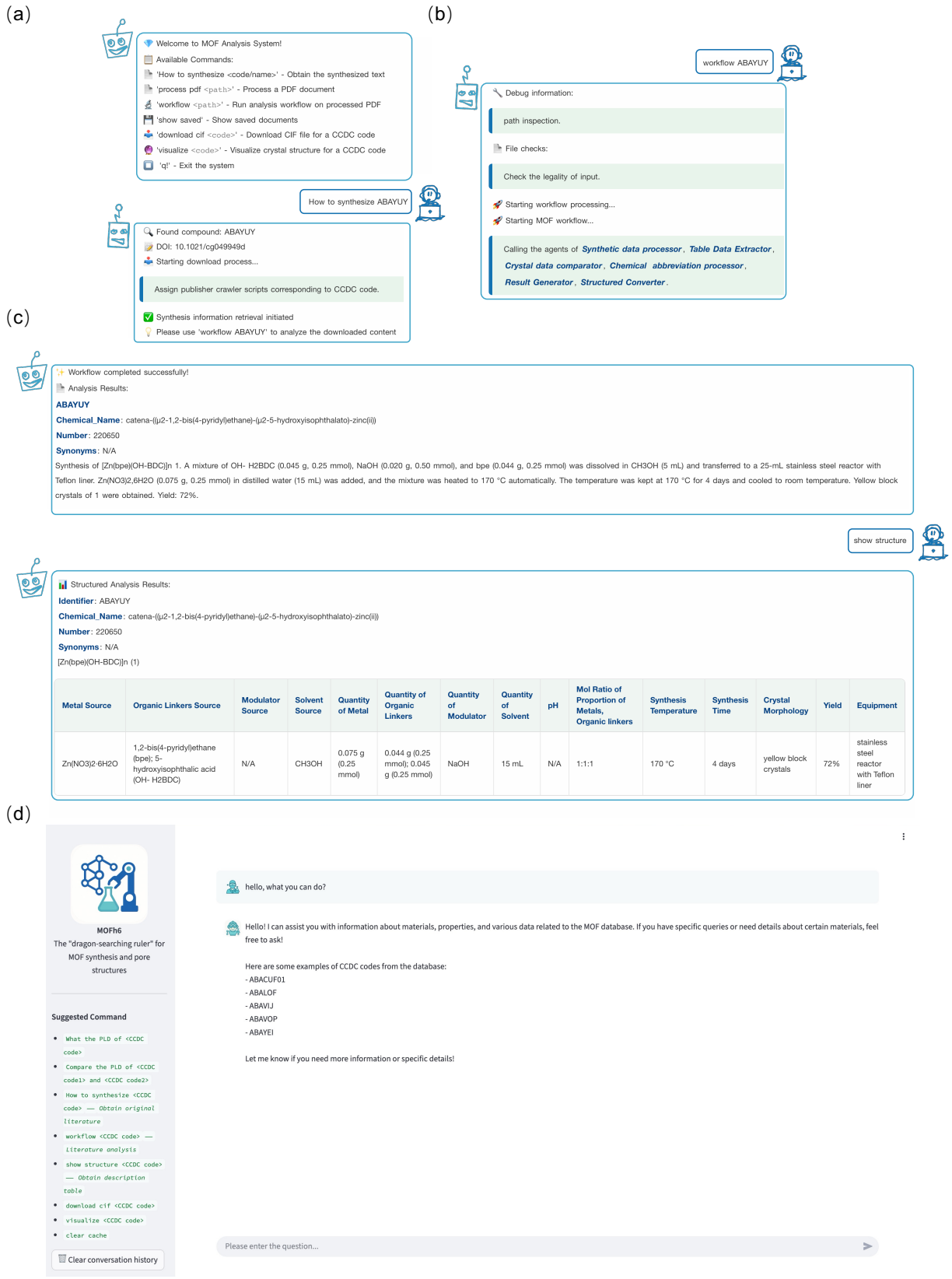

Figure 3: Query & Answer of MOFh6. User commands for initiating literature crawling (a), sequential agent-based mining of synthesis data (b), output of structured synthesis parameters (c).

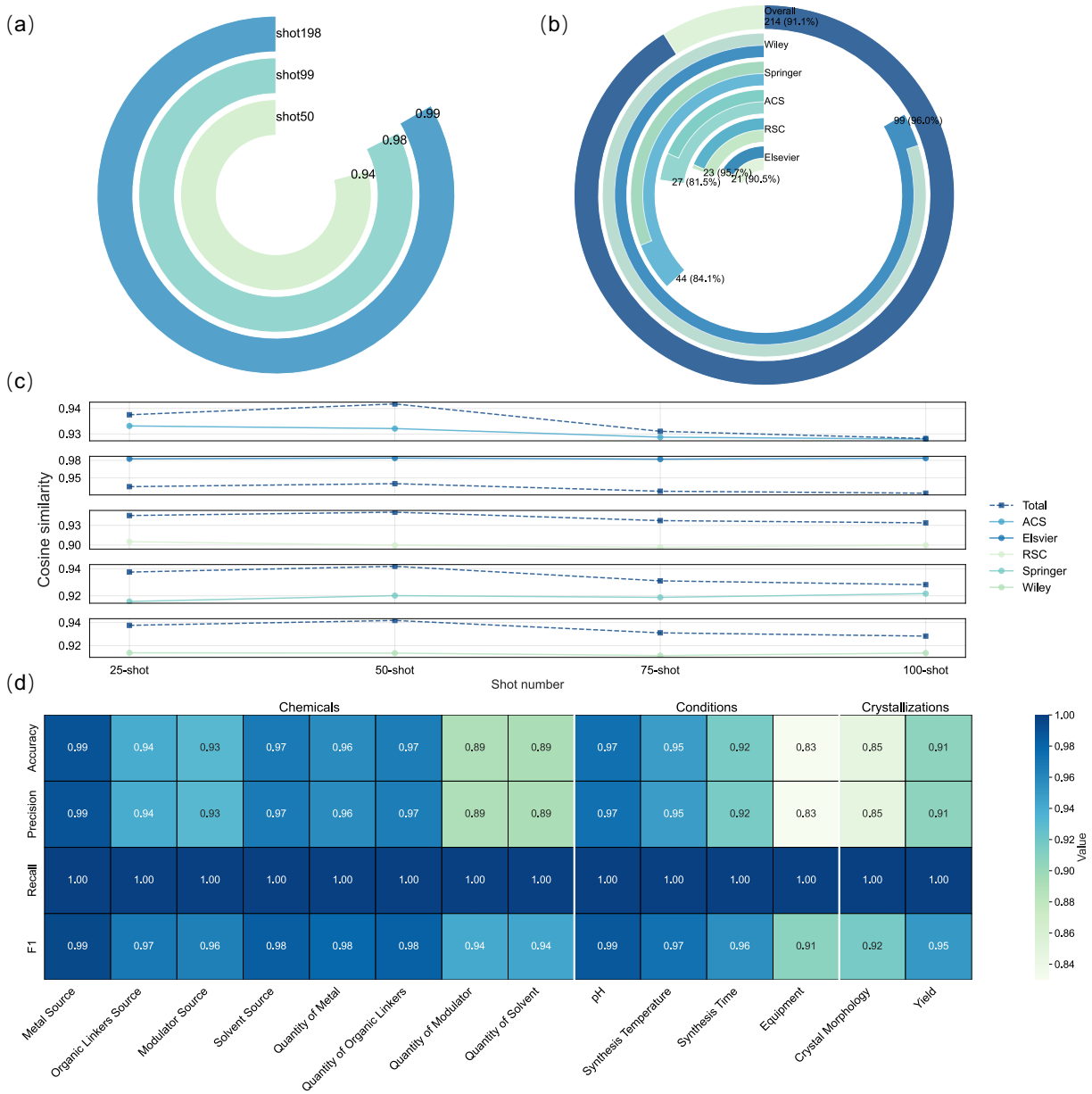

Figure 4: Agents performance evaluation: The performance of different sample fine-tuning models on their test set in the dataset (a), the ability of coreference resolution (b), the ability of the model to extract specific MOFs synthesis paragraphs under different pool sizes (c), and the ability of the model to extract structured data (d).

Supporting Information

Reshaping MOFs text mining with a dynamic multi-agents framework of large language model

Table of Contents

Text S1 Agents

I. Crawler module

The literature crawler was built on requests and BeautifulSoup to construct an HTML parsing engine, which simulates browser behavior through customized request headers to circumvent anti-crawling mechanisms. A parallel request management and frequency control strategy was adopted to improve collection efficiency, combined with publisher APIs to directly obtain PDF/HTML content. Supporting materials were automatically downloaded through Selenium executing XPath pattern matching, with intelligent processing performed for different formats: ZIP packages were automatically uncompressed, while doc and docx documents were uniformly converted to PDF through the invocation of open-source software LibreOffice [†]. Finally, all content was converted to standardized TXT text via PyMuPDF, with result feedback and log tracking mechanisms integrated to ensure system maintainability (Figure S1).

II. Synthetic data parsing agent

A supervisedly fine-tuned GPT-4o-mini model was adopted as the synthesis data parsing agent, with its training data construction following strict annotation standards: the COO, played by human experts, constructed a benchmark annotation dataset focusing on precise identification of MOF synthesis paragraphs, established a standardized annotation framework, and excluded organic ligand synthesis interference items. Complete parsing of bridging reference contexts was performed with recording of 13 key parameter evolution paths including metal salts, organic ligands, additives, and solvent systems (Table S2), with recalculation conducted if there were quantity changes. Filtering of post-synthesis characterization data was simultaneously executed, including elemental analysis (EA), spectral data, nuclear magnetic resonance (NMRs), and Fourier-transform infrared spectroscopy (FT-IR) that were not synthesis-related (Figure S12, with shot examples shown in Figure S36). The model employed a dynamic learning strategy: an initial learning rate of 5×10^{-5}, adaptive adjustment based on validation set loss values, a maximum training period of 4 epochs, and a batch processing size of 8 (Figure S2).

III. Table data parsing agent

The table data parsing agent was constructed based on a GPT-4o-mini model driven by few-shot prompt engineering, specifically designed to process heterogeneous data sources (Figure S9). The table data parsing agent can both parse formatted data in discrete tables (5-shot) and capture implicit crystallographic descriptions in text paragraphs (3-shot) (Figure S17-Figure S24). Through customized prompt templates, the system locks onto compound names, empirical molecular formulas, molecular weights, crystal systems, space groups, lattice constants (a, b, c), unit cell angles (α, β, γ), and crystal colors. For data heterogeneity challenges, a dynamic synonym mapping mechanism was developed. In terms of data integrity control, a dual-threshold filtering strategy was set, with an automatic entry elimination mechanism triggered when any two key parameters were missing, ensuring the completeness and usability of output data (Figure S3).

IV. Crystal data comparison agent

The crystal data matching system was constructed based on 5-shot prompt engineering with GPT-4o, employing a dual-level short-circuit evaluation mechanism to achieve precise mapping of structures between the CCDC database and literature descriptions (Figure S10, shot examples as shown in Figure S25-Figure S29). The primary comparison focuses on lattice parameters, setting a joint tolerance threshold (90% match degree) for crystal systems, space groups, unit cell lengths, and angles. If the threshold is not reached, a secondary chemical composition verification layer is activated, mandating complete matching of metal elements and a total chemical formula similarity of $\geq$30%. The

[†]https://www.libreoffice.org

system integrates a cleaning engine that automatically converts space group symbols, eliminates numerical format interference, performs unit conversions, and ensures standardization and comparability of comparison data (Figure S4).

V. Chemical abbreviation resolution agent

The chemical abbreviation resolution agent was used to solve the co-reference resolution problem of ligand naming such as H_xL_x, L_xH_x, L_x ($x \geq 0$) in literature (Figure S11). This processor is based on a library of regular expression relationships for 15 types of full name-abbreviation mappings, combined with GPT-4o's dynamic context-aware mechanism to achieve precise resolution. To suppress model hallucinations, a prompt framework was constructed containing 15-shot effective mapping samples and 4-shot ambiguity samples (shot examples as shown in Figure S30-Figure S35), while adopting a triple filtering mechanism: pattern compliance verification [24], automatic abandonment of full names containing metal elements, and verification of abbreviation-full name pair co-occurrence. The resolution process sequentially includes document structure parsing, cross-paragraph semantic association analysis, mapping relationship probability verification, and standardized output organization (Figure S5).

VI. Post processor

The post-processor is responsible for fine-grained processing and organizational management of integrated outputs, ensuring reliable transmission of processing results to downstream nodes through the StateSchema mechanism, while possessing comprehensive error handling capabilities to maintain workflow continuity. This module employs a document splitting engine, which uses double blank line paragraph separator recognition technology combined with first-line unique identifiers, automatically extracting CCDC codes to precisely split integrated documents into independent compound files while preserving the original semantic hierarchical structure. To enhance data traceability, standardized processing reports are simultaneously generated, recording timestamps, total number of processed files, and detailed file lists (Figure S6).

VII. Result generator agent

The result generator agent is the core integration node of the MOF data processing system, responsible for fusing multi-source heterogeneous data into unified structured outputs. It adds processed file paths to the StateSchema object, ensuring reliable data transmission in the workflow. Through a three-way data integration strategy, information flows from the synthesis data parsing agent, crystal data comparison agent, and chemical abbreviation resolution agent are coordinated. Its core method is a fusion mechanism with compound identifiers as primary keys, traversing each target compound, extracting its structural parameters, synthesis method descriptions, and related abbreviation explanations, and applying unified formatting templates to generate standardized entries (Figure S7).

IX. Structured conversion agent

The structured conversion agent complexly extracts key chemicals and quantities from unstructured synthesis information, possessing data parsing modules and information extraction and structuring modules (Figure S8). The system first performs preprocessing and parsing of the original text, then uses 4-shot GPT-4o-mini for information extraction (Figure S13-Figure S16), ultimately generating standardized structured data tables.

Text S2 Hallucination relief

Previous research has demonstrated that endowing LLMs with carefully designed prompt engineering strategies significantly enhances their reliability and calibration in complex chemical tasks [23, 37, 57]. Figure S8-Figure S12 are the few-shot prompt templates designed by us, with core designs including the introduction of Chain-of-Thought mechanisms, guiding models to explicitly generate intermediate logical nodes through step-by-step reasoning frameworks, enhancing the transparency of LLM multi-step description parsing [58, 59]. Additionally, LLMs are required to use JSON structured templates for content transmission throughout the running process, leveraging their strict semantic hierarchy and type constraint characteristics to reduce ambiguity and vagueness in prompts, enabling models to more accurately understand user intentions and needs [60]. Figure S13-Figure S36 are the carefully designed shot examples in these prompts.

Text S3 Agents' performance evaluation

For the special nature of chemical semantics, BERT-based (eq. S1) BiomedBERT and PubMedBERT models were used to evaluate specialized chemical texts. When comparing sentences between manually annotated and model-generated paragraphs, a comprehensive text preprocessing mechanism was adopted in this thesis. This mechanism recognizes and processes various patterns unique to chemical literature through regular expressions, including removing synthesis titles like "Synthesis of XXX:", eliminating elemental analysis (EA) and FT-IR spectral data, NMRs data, and standardizing temperature and time expressions (unifying "100 °C" and "100 oC" as "100 C", converting "24 hours" to "24h"). These preprocessing rules were specifically designed for non-standard expressions in chemical synthesis descriptions, endowing machines with human-like subjectivity, as this content does not affect the final judgment of synthesis descriptions.

Secondly, given that synthesis methods are described in the main text of scientific papers, the PubMedBERT model[‡] [61] specifically developed for the biomedical field was used for text representation. This model can effectively capture semantic features in chemical synthesis texts, generating high-dimensional vector representations. For each pair of synthesis methods matched by CCDC code, the system first checks whether the preprocessed texts are completely identical; if identical, the highest similarity (1) is directly assigned; if not completely identical, the similarity is quantitatively assessed by calculating the cosine similarity between the embedding vectors of the two text segments (Equation 5-1). Finally, the system generates detailed comparison results, including the original text, preprocessed text, embedding similarity scores, and identification of complete matches for each matching record. The overall dataset's average similarity was also calculated, with results arranged in descending order of similarity to facilitate priority review of the most similar synthesis method pairs.

When comparing the content within cells of data frames, an optimized rule matching engine was first adopted by the system. This engine captures expression patterns unique to the chemical domain through multiple specialized rules, including: percentage expression recognition (such as equivalence matching between "35%" and "0.35"), abbreviation mapping within parentheses, chemical formula pattern matching, text normalization processing (removing spaces and Unicode variant characters), yield description matching, equipment keyword recognition (such as "Teflon lined autoclave"), substance amount and mass description pairing (such as "0.25 mmol, 0.061 g" and "0.061 g (0.25 mmol)"), and solvent volume accumulation matching. These rules are designed for specific expression habits in chemical literature and can efficiently identify semantically equivalent but differently expressed descriptions. For each pair of synthesis method cells matched by CCDC code, the system first attempts to determine content equivalence through the rule engine; if the rule judgment is uncertain, an appropriate semantic model is selected based on the nature of the cell content, precisely quantifying the semantic proximity between different expressions by calculating the cosine similarity of text vectors. This hierarchical, multi-model similarity assessment strategy significantly improved the system's accuracy in determining the equivalence of chemical synthesis method descriptions.

Secondly, for situations where rule matching cannot determine, a dual semantic model strategy was employed. Different pre-trained models were selectively applied for different types of synthesis information. Considering that chemicals appear simultaneously in abstracts and main text, the evaluation of chemical sources used the BiomedBERT model[§] [56] trained on both abstracts and full texts for cell content representation, distinct from the sentence-level PubMedBERT model, while the Sentence Transformer model[¶] was applied for synthesis condition information to capture general semantic relationships.

$$E(x) = \mathrm{Normalize}\left(\frac{\sum_{i=1}^{n} f(x_i)\, M_i}{\sum_{i=1}^{n} M_i}\right) \tag{S1}$$

[‡]https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext

[§]https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

[¶]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Where $E(x)$ denotes the final embedding vector of text x, $f(x_i)$ is the embedding of the i-th token in x extracted by transformer models (e.g. BiomedBERT, PubMedBERT, MiniLM), M_i is the mask value for token i (0 for padding, 1 for actual tokens), n is the token sequence length, and $\mathrm{Normalize}$ denotes the L2-normalization operation.

Text S4 Query & Answer implementation principle

MOFh6 converts users' unstructured queries into JSON structured data processable by the system through LLMs. When a user issues a query (Figure S46), the system first analyzes and clarifies which specific query type (query_type) it belongs to, such as property query (property), range query (range), comparison analysis (comparison), statistical calculation (statistical), reset (reset), greeting (greeting) or general chat (chat). For example, when a user asks "What is the PLD of MOF-5?", the system identifies it as a property query and sets the query_type to `"property"` (Figure S38).

The system's context management capability is clearly reflected through the uses_context field. When the system detects that the current query depends on previous conversation content, the uses_context field is set to true. For example, after a user asks "What is the PLD of VUJBEI?", if they continue to ask "What about its density?", the system can recognize this query's reference to the previous material and understand the current query 's context by referring to the last_materials field (e.g., `["VUJBEI"]`) (Figure S39 and Figure S40). After each query is successfully executed, the system automatically updates the last_query, last_materials, last_properties, and last_result fields to facilitate smoother subsequent context-dependent queries.

During parsing, the system simultaneously extracts material identifiers involved in the query and stores them in the materials array. For instance, in the user query "Compare the density of VUJBEI and QOWTIG", the system records `["VUJBEI", "QOWTIG"]` in the materials field. Additionally, the system can accurately identify the material properties of interest to the user, such as "PLD", "density", or "surface area", and after normalization, saves these properties in the properties array, such as `["PLD (ÃĚ)"]` or `["Density (g/cm3)", "Accessible\_Surface\_Area (m2/cm3)"]` (Figure S40).

Range query is an important function of MOFh6, with the system processing range constraints involving multiple properties through the range object. For example, when a user asks "Find MOFs with PLD between 7.5 and 10 Å and LCD between 10 and 16 Å", the system automatically generates a structure containing multi-property constraints: `{"min": {"PLD": 7.5, "LCD": 10}, "max": {"PLD": 10, "LCD": 16}}`. For more complex multi-condition range queries, such as "Give me MOFs with PLD between 7.5-10 Å, LCD between 10-16 Å, and VSA between 2000-2400 m^2/cm^3", MOFh6 can accurately parse this into `"min": {"PLD ÃĚ": 7.5, "LCD ÃĚ": 10, "VSA m2/cm3": 2000}`", `""max": {"PLD ÃĚ": 10, "LCD ÃĚ": 16, "VSA m2/cm3": 2400}` (Figure S41 and Figure S44).

For queries involving statistical or comparative operations, the system clearly records relevant operation information in the operation object. For example, when a user asks "What is the average density of the MOF-5 series?", the operation field will contain `"type": "mean", "value": null` (Figure S43); when requesting "Find the MOF with the maximum density", it is recorded as `"type": "max", "value": null`. The system's reasoning steps for parsing queries are also recorded in the reasoning array, clearly showing the logical pathway of how the system understands query intentions, thereby providing decision transparency (Figure S42).

For large amounts of results requiring paged display, the system manages through page_size and paged_index fields. When a user requests "Show more results" or "Give me 5 more", the system identifies this as a paging query (query_type as `"paging"`) and provides the next batch of results based on page_size (such as 5 or 10) and the current paged_index (such as 0, 5, or 10) (Figure S44).

Additionally, the system maintains complete query history records (query_history), recording previous questions, parsing results, and involved materials and properties. This not only supports deep understanding of context but also provides a foundation for system learning and improvement. When a user's query is relatively vague, the system can also reference parsing results of similar queries in history, thereby enhancing the accuracy of current parsing (Figure S42). Finally, the system generates natural, fluent, and content-rich responses based on parsing results and retrieved data through the _final_decision_with_llm.

Text S5 Cost account

MOFh6, through its multi-agent collaborative architecture, deeply integrates LLM's natural language processing capabilities with rule engines, achieving the dual optimization of intelligent task decomposition and dynamic result aggregation. This design significantly reduces the hallucination risks common in traditional LLM applications while demonstrating outstanding advantages in computational efficiency and economy. In synthesis information extraction tasks, the system completes semantically complete synthesis description paragraph extraction for a single PDF document in just 9.6 seconds (Figure S51a), far lower than the 69.2 seconds for single PDF processing by ChatGPTChemistryAssistant proposed by Zheng et al. [23]. In the structured description confirmation stage, the system's average time for precisely locating specific MOFs and their structured descriptions is controlled within 36 seconds (Figure S51b). In terms of economy, MOFh6's data extraction cost for processing 100 papers is only 4.24$ (Figure S51c), achieving about 76% cost savings compared to the L2M3 system (17.94$/100 papers) developed by Kang et al. [25], with the cost per 100 papers for single agent processing also below 2.5$ (Figure S51d). This significant improvement in efficiency and cost is mainly attributed to MOFh6's precise response mechanism under rule engine constraints and the efficient collaboration mode between distributed agent networks.

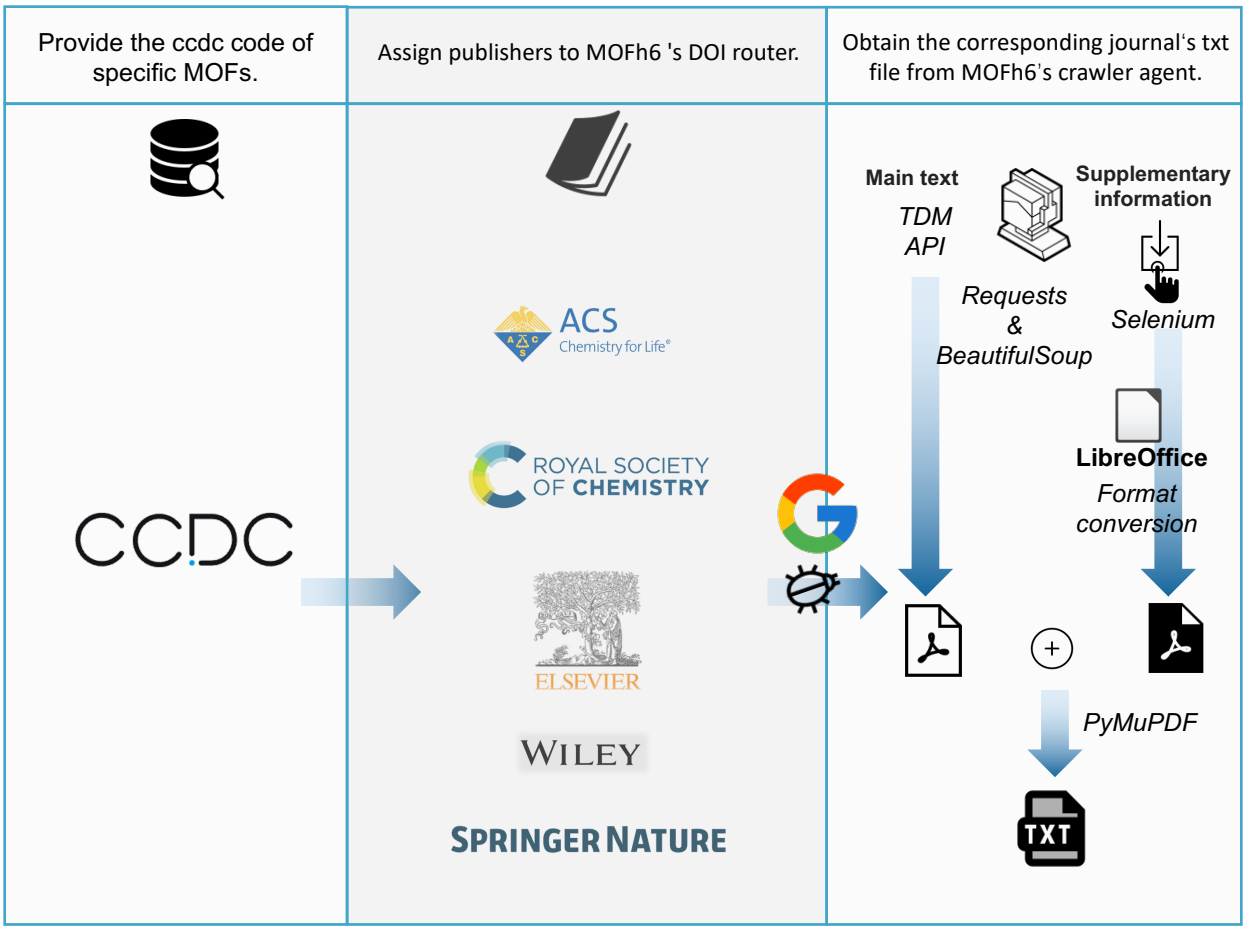

Figure S1: The crawler module of MOFh6

Figure S2: Processing logic of synthetic data parsing agent

	1	2	3
empirical formula	$C_{16}H_{11}DyN_2O_5$	$C_{16}H_{11}N_2O_5Yb$	$C_{16}H_{11}GdN_2O_5$
F_w	473.77	484.31	468.52
cryst size [mm]	0.18 × 0.17 × 0.15	0.18 × 0.17 × 0.15	0.18 × 0.17 × 0.16
cryst syst	tetragonal	tetragonal	tetragonal
space group	$I\bar{4}$	$I\bar{4}$	$I\bar{4}$
a [Å]	18.9363(12)	18.885(7)	19.0364(14)
b [Å]	18.9363(12)	18.885(7)	19.0364(14)
c [Å]	8.2997(10)	8.222(3)	8.3437(12)
V [Å^3]	2976.1(4)	2932.3(19)	3023.6(5)
Z, D_c [g cm^{-3}]	8, 2.115	8, 2.194	8, 2.058
$h/k/l$	−22, 16/−22, 20/− 9, 9	−22, 16/−21, 22/− 9, 9	−22, 22/−22, 22/− 6, 9
$F(000)$	1816	1848	1800
μ [mm^{-1}]	5.052	6.409	4.417
reflections collected/unique	8693/2634	7742/2594	7792/2667
R_{int}	0.0435	0.0763	0.0168
data/restraints/params	2634/0/217	2594/127/217	2667/1/217
R_1,[a] wR_2,[b] [$I > 2\sigma(I)$]	0.0235, 0.0440	0.0357, 0.0897	0.0126, 0.0274
R_1, wR_2 [all data]	0.0246, 0.0445	0.0363, 0.0903	0.0132, 0.0276
max. and min transmission	0.7226 and 0.6608	0.4465 and 0.3917	0.5384 and 0.5036
GOF on F^2	1.050	1.082	1.029
$\Delta\rho_{max}$, $\Delta\rho_{min}$ [e·Å^{-3}]	0.557, − 0.515	1.937, − 2.714	0.293, − 0.290

[a] $R_1 = \Sigma(||F_0| - |F_c||)/\Sigma|F_0|$; [b] $wR_2 = [\Sigma w(|F_0|^2 - |F_c|^2)^2/\Sigma w(F_0^2)^2]^{1/2}$.

Inorg. Chem. 2015, 54, 1, 153–160


```
{
    "Complexes": "1",
    "Empirical formula": "C16H11DyN2O5",
    "Formula weight": "473.77",
    "Crystal system": "tetragonal",
    "Space group": "I4̄",
    "a (Å)": "18.9363(12)",
    "b (Å)": "18.9363(12)",
    "c (Å)": "8.2997(10)",
    "alpha (°)": "N/A",
    "beta (°)": "N/A",
    "gamma (°)": "N/A",
    "Crystal size (mm)": "0.18 × 0.17 × 0.15"
},
```

```
{
    "Complexes": "2",
    "Empirical formula": "C16H11N2O5Yb",
    "Formula weight": "484.31",
    "Crystal system": "tetragonal",
    "Space group": "I4̄",
    "a (Å)": "18.885(7)",
    "b (Å)": "18.885(7)",
    "c (Å)": "8.222(3)",
    "alpha (°)": "N/A",
    "beta (°)": "N/A",
    "gamma (°)": "N/A",
    "Crystal size (mm)": "0.18 × 0.17 × 0.15"
}
```

```
{
    "Complexes": "3",
    "Empirical formula": "C16H11GdN2O5",
    "Formula weight": "468.52",
    "Crystal system": "tetragonal",
    "Space group": "I4̄",
    "a (Å)": "19.0364(14)",
    "b (Å)": "19.0364(14)",
    "c (Å)": "8.3437(12)",
    "alpha (°)": "N/A",
    "beta (°)": "N/A",
    "gamma (°)": "N/A",
    "Crystal size (mm)": "0.18 × 0.17 × 0.16"
}
```


Figure S3: Extraction logic of table data parsing agent

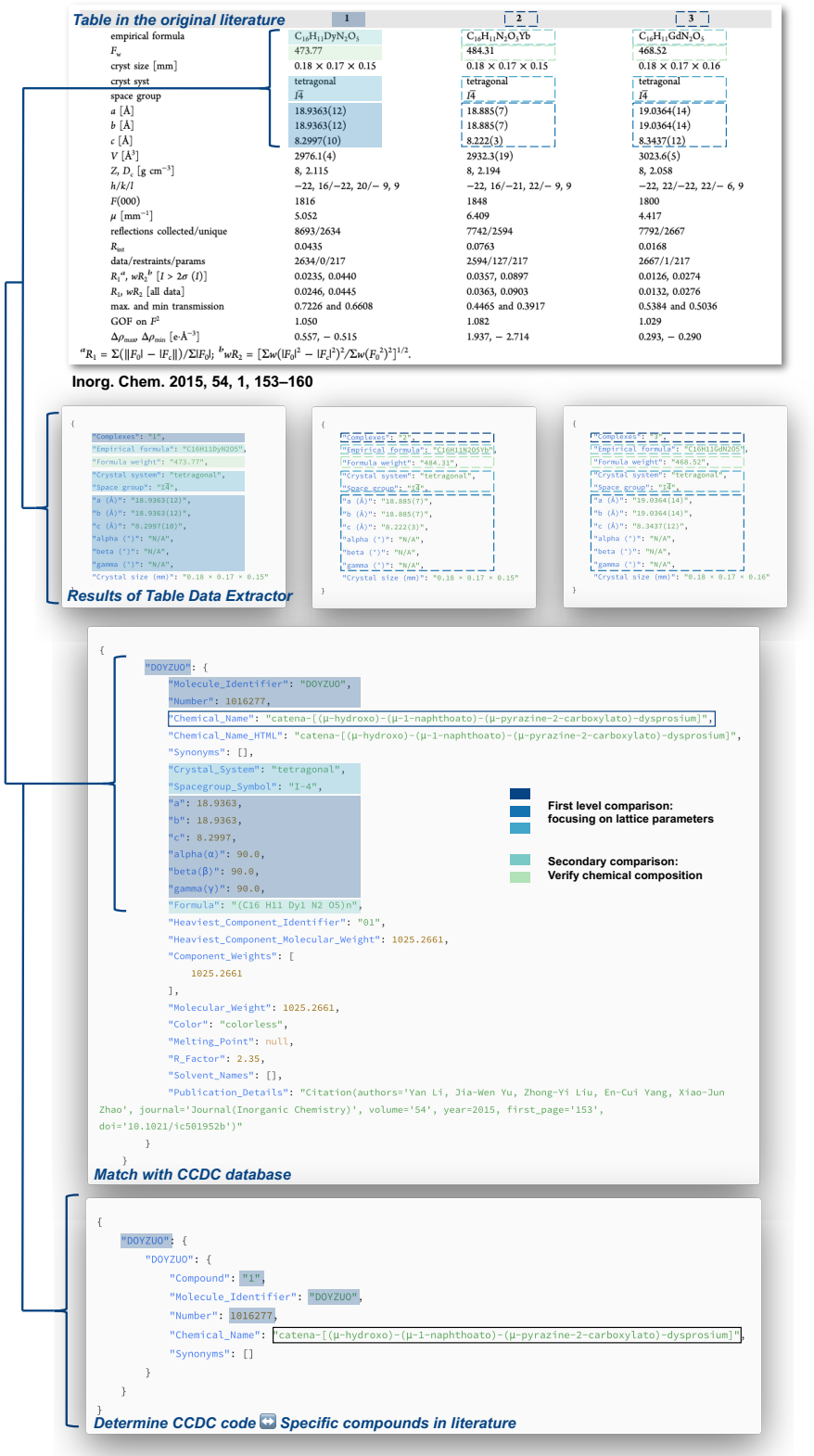

Figure S4: Comparison logic of crystal data comparison agent

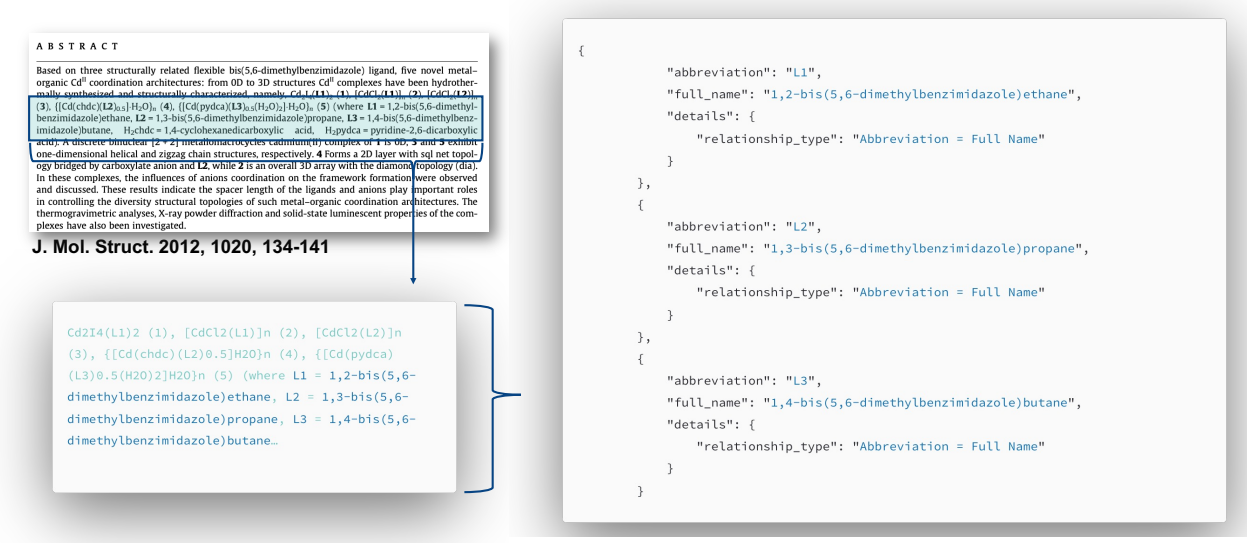

Figure S5: Coreference resolution case

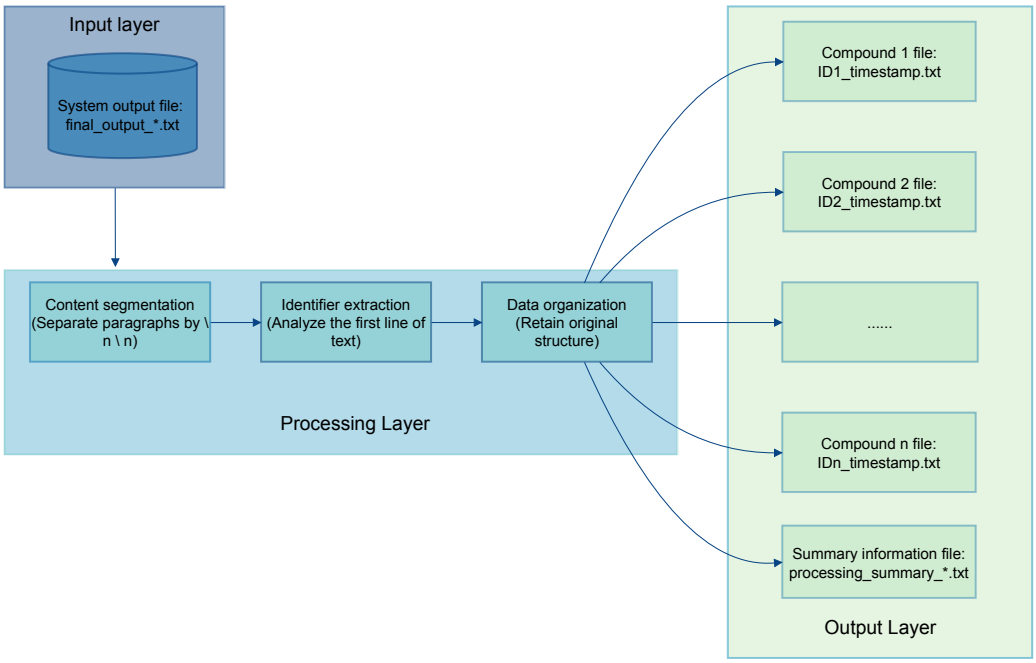

Figure S6: Logic of post processor

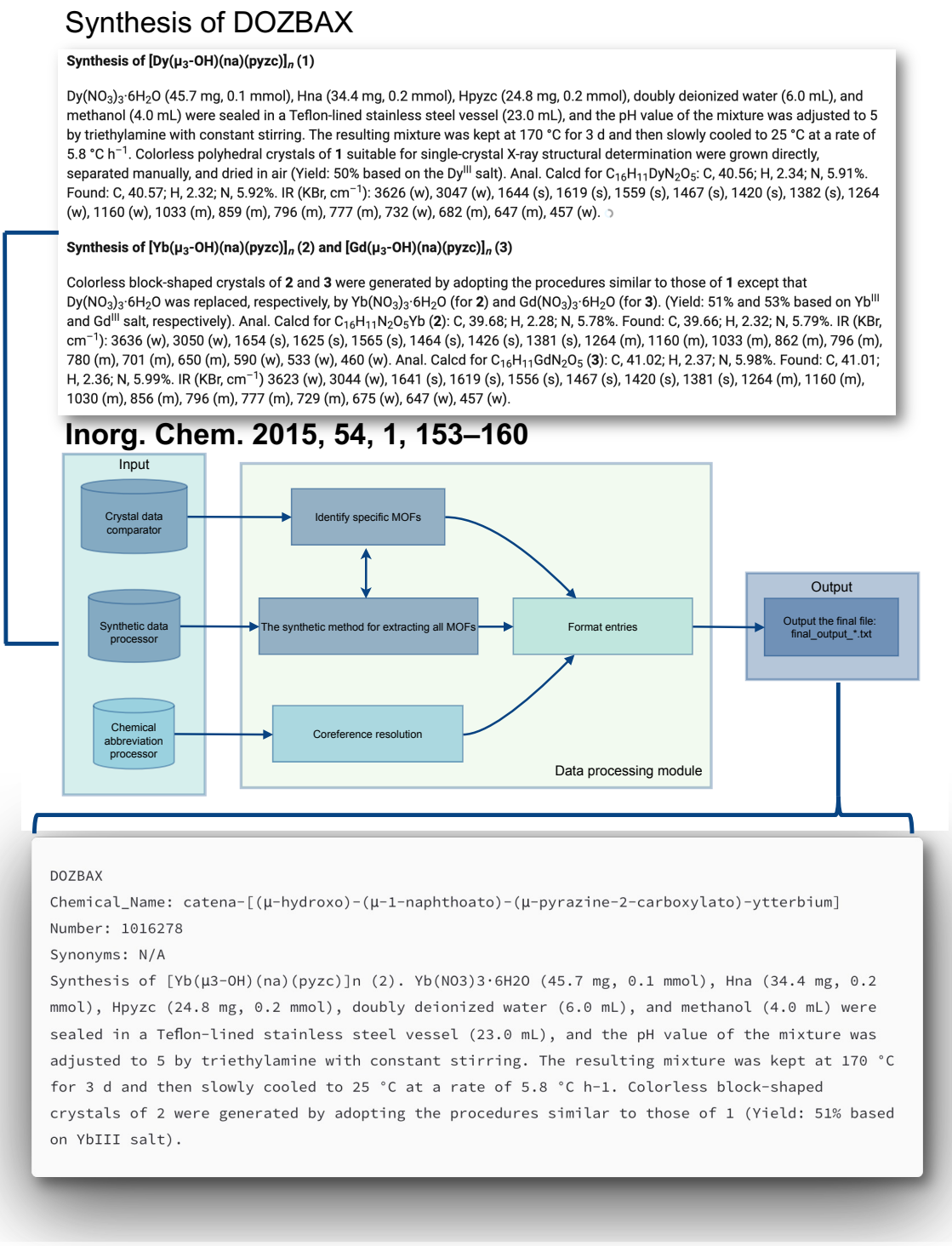

Figure S7: Logic of result generator agent

- **Role**: As a MOF synthesis expert, you are responsible for analyzing the input information, extracting key information about synthesises, and outputting them in the structured markdown table.
- **Input Information**:

  Identifier: {identifier}
  Chemical_Name: {chemical_name}
  Number: {number}
  Synonyms: {synonyms}
  Content: {content}
  Abbreviation: {Abbreviation}
  Full_Name: {full_name}
- **Step**
  1.Extract the **Identifier**, **Chemical_Name**, **Number**, and **Synonyms** of the input file, and start standardizing the **Content**.
  2.Identify **Content** related to synthesis, typically involving multiple chemical substances under certain conditions.
  3.Fill the extracted key information into the corresponding position in the markdown table.
  4.Output the markdown table.
  - Output Information:
    Identifier: {identifier}
    Chemical_Name: {chemical_name}
    Number: {number}
    Synonyms: {synonyms}
    {Compound_Name_or_Number_from_Synthesis}

Metal Source	Organic Linkers Source	Modulator Source	Solvent Source
[Extract from content]	[Extract from content]	[Extract from content]	[Extract from content]

Quantity of Metal	Quantity of Organic Linkers	Quantity of Modulator	Quantity of Solvent
[Extract from content]	[Extract from content]	[Extract from content]	[Extract from content]

pH	Synthesis Temperature	Synthesis Time	Equipment
[Extract from content]	[Extract from content]	[Extract from content]	[Extract from content]

Crystal Morphology	Yield
[Extract from content]	[Extract from content]

```
**Examples**:
  - 3-shots
```

5.If the input file contains **Abbreviation** (usually used to describe Organic Linkers), **Full Name** (usually used to describe Organic Linkers), the final markdown table needs to replace the original **Abbreviation** in the text with **Full Name**.

```
**Example**:
  - 1-shot
```

- **Rules**:

1. Only extract information explicitly stated in the content
2. Use "N/A" for any missing information
3. Present the data in exact format shown below

4. Do not include any additional explanatory text
5. Ensure all extracted values maintain their original units and notations

- **Important Notes**:
  - Extract compound name/number exactly as written in synthesis text
  - Keep all original units and measurements
  - Maintain exact chemical formulas and notations
  - A markedown cell allows filling with multiple substances
  - When multiple substances appear in the same markedown cell, use the ; to separate them
  - Multiple organic linkers may exist during a synthesis process
  - Modulators are usually substances that can regulate pH, such as inorganic acids, inorganic bases, and triethylamine, etc.
  - Only include information explicitly stated in the content
  - Do not add any interpretations or assumptions
  - Do not include any examples or additional text

Figure S8: Prompt of structured conversion agent.

- **Role**: Data Analysis Expert and Crystallographic Data Organizer
- **Background**: The user needs to extract specific crystallographic data from text paragraphs containing the keywords **"Table"** or **"Crystal data."** This information may appear in two formats: as structured tables or embedded within textual references.
- **Profile**: You are a professional crystallographic data analyst skilled in extracting and organizing key information from complex datasets. You can accurately identify and process data related to crystal structures and properties.
- **Skills**: Expertise in crystallography, data extraction, and information organization. You can quickly identify and extract essential crystallographic data from text.
- **Goals**: Accurately extract relevant data associated with the following keywords from text paragraphs containing **"Table"** or **"Crystal data"**:

Keywords		
Complexes	Space group	alpha (α)
Empirical formula	a (Å)	beta (β)
Formula weight	b (Å)	gamma (γ)
Crystal system	c (Å)	Colour

- **Important Notes**:
  - **Synonyms**: Be aware that the keywords may have synonyms; do not overlook their presence.
- **Constraints**:
  - The extracted data must be accurate and presented in an easily understandable dictionary format, ensuring completeness and correctness.
  - Only extract data when all specified key values are present. If two or more key values are missing (e.g., marked as N/A), ignore the entire entry.
  - Do not quote or include content from the **Examples** in the results.
- **Workflow**:
  - Let's proceed step by step with the operation.
  1. Learn the content of the example, which provides the text should be recognized, as well as the extracted content and form.

```
**Examples**:
**Contents Table**:
 - 5-shots
**Contents Crystal data**:
 - 3-shots
```

  2. Identify and locate text paragraphs containing **"Table Crystal data.../Crystallographic data..."** or **"Crystal data."**
  3. Within these paragraphs, search for and extract data associated with the specified keywords.
  4. Organize the extracted data into a structured dictionary format.
- **Output Format**:
  - A structured dictionary that clearly displays each keyword and its corresponding data.

Please process the following content and return it in JSON format:

{{file_content}}

And provide the output in the Output Format.

Figure S9: Prompt of table data parsing agent

- **Role**: Data Comparison and Analysis Expert
- **Background**: The user requires a comparison of specific data between two JSON files and the output of results based on specific logical criteria.
- **Profile**: You are a professional data analyst with expertise in handling and comparing complex datasets. You excel at identifying similarities and differences between data elements with precision.
- **Skills**: You possess advanced data processing capabilities, enabling you to extract key information from JSON files and apply logical evaluation to assess data consistency.
- **Goals**: Extract specific data from two JSON files, compare the data based on the given logical criteria, and output results that meet the specified conditions.
- **Constraints**: The comparison and analysis must be accurate, with clear, precise output that adheres to the user's specified format.
- **OutputFormat**: JSON format containing the required tags and data fields.
- **Workflow**:
  - Let's proceed step by step with the operation.

  1. **Learning Examples**: Here are some examples for your reference, these examples are for educational purposes only.

```
**Examples**
 - 5-shots
```

  2.Process the following documents:

  - **a.json**: {{a_json}}
  - **b.json**: {{b_json}}

  3. Extract the following fields from `a.json`: `"Crystal_System"`, `"Spacegroup_Symbol"`, `"a"`, `"b"`, `"c"`, `"alpha(α)"`, `"beta(β)"`, `"gamma(γ)"`, `"Formula"`, `"Component_Weights"`.
  4. Extract the following fields from `b.json`:
     - `"Crystal system"`, `"Space group"`, `"a (Å)"`, `"a (nm)"`, `"b (Å)"`, `"b (nm)"`, `"c (Å)"`, `"c (nm)"`, `"alpha (°)"`, `"beta (°)"`, `"gamma (°)"`, `"Empirical formula"`, `"Formula weight"`.
  5. Now, proceed with the grouping:
     - **Group 1**: Compare the following pairs:
       - `"Crystal_System"` - `"Crystal system"`
       - `"Spacegroup_Symbol"` - `"Space group"`
       - `"a"` - `"a (Å)"`, `"b"` - `"b (Å)"`, `"c"` - `"c (Å)"`
       - `"alpha(α)"` - `"alpha (°)"`, `"beta(β)"` - `"beta (°)"`, `"gamma(γ)"` - `"gamma (°)"`
     - **Group 2**: Compare the following pairs:
       - `"Formula"` - `"Empirical formula"`
       - `"Component_Weights"` - `"Formula weight"`
  6. Unit Conversion Rule:
     - The units for `"a"`, `"b"`, `"c"` may be in Å (angstroms) or nm (nanometers).
     - **Conversion Factor**: 1 nm = 10 Å.
     - When comparing these values between the two JSONs, convert the units if necessary so that the values are in the same unit before comparison.
     - For example, if one value is in nm and the other is in Å, convert the nm value to Å by multiplying by 10.
  7. Compare the similarity of the data in **Group 1**:

- Ensure that the similarity is greater than **90%**.
- If any floating-point numbers are enclosed within `" "`, ignore the quotation marks and treat them as numbers.
- Only compare the keys that are present. If a key is missing (typically in `b.json`), skip that comparison.
- **Apply the Unit Conversion Rule** for `"a"`, `"b"`, `"c"` as specified in step 6.

8. Compare the similarity of the data in **Group 2**:
   - Ensure that the similarity is greater than **30%**.
   - The most importance in Group 2 is to ensure that the metal elements in the two objects to be compared are completely identical.
9. If the conditions for **Group 1** are satisfied, skip the comparison for **Group 2**.
10. If the data meets the similarity requirements, output the result in the specified format:

- **Output Format**:

```
{
  "<Molecule Identifier>":
    {
       "Compound": "<Compound>",
       "Molecule_Identifier": "<Molecule Identifier>",
       "Number": "<Number>",
       "Chemical_Name": "<Chemical Name>",
       "Synonyms": "<Synonyms>"
    }
}
```

- **Important Notice**:
  - Please strictly follow the output format in step 10 and only output JSON results.
  - Please ensure that the output JSON is **valid** and can be parsed by a standard JSON parser.
  - **All keys and values (especially string values) must be enclosed in double quotation marks.**
  - Do not include any explanation, reasoning, or additional text.
  - Provide the final JSON result directly.

Figure S10: Prompt of crystal data comparison agent

Role: You are a Senior Text Analysis Expert and Terminology Recognition Consultant with extensive knowledge in organic chemistry.
Your primary role is to
accurately extract chemical abbreviations that strictly match the following regular expression patterns:

1. **HxLx**: `H\d*L\d*`
   - Matches abbreviations starting with 'H', followed by zero or more digits, then 'L', followed by zero or more digits.
   - Examples: HL, H2L, H3L4.
2. **Lx**: `L\d+`
   - Matches abbreviations starting with 'L', followed by one or more digits.
   - Examples: L1, L23.
3. **LxHx**: `L\d+H\d+`
   - Matches abbreviations starting with 'L', followed by one or more digits, then 'H', followed by one or more digits.
   - Examples: L1H1, L2H3.

Important Notes:

- **Abbreviations must strictly conform to these patterns**. Do not extract abbreviations that do not match these patterns.
- **The abbreviations should not exceed three tokens**.

Additionally, you will identify their corresponding full chemical names, ensuring that metal elements from the periodic table do not appear in the full names.

You must understand and recognize specific positional relationships between abbreviations and full names, including the use of equal signs (`=`), colons (`:`), and parentheses (`()`). The relationships to identify are:

Abbreviation = Full Name	**Full Name = Abbreviation**	**Abbreviation : Full Name**
Full Name : Abbreviation	**Full Name (Abbreviation)**	**Abbreviation (Full Name)**
Abbreviation (Full Name, with Quantitative Dose)	**Full Name (Abbreviation, with Quantitative Dose)**	**Full Name Abbreviation**
Abbreviation Full Name	**Abbreviation is/are/was/were Full Name**	**Full Name is/are/was/were Abbreviation**
Abbreviation, Full Name	**Full Name, Abbreviation**	

- **Exceptional case:**

Abbreviation+connector (e.g., “respectively,...”、“in that order,...”、“corresponding to,...”, “sequentially,...”, etc)+Full Name

Tasks:

- Let's proceed step by step with the operation.

1. **Learning Examples**: The following positive and negative examples are provided to illustrate correct and incorrect extractions. These examples are for educational purposes only.

```
**Examples**:
**Positiv Examples**:
 - 15-shots
***Negative Examples**:
 - 4-shots
```

2. **Traverse the Document**: Examine the entire content of the following document to identify all abbreviations that strictly match the specified regular expression patterns (`H\d*L\d*` , `L\d+` , `L\d+H\d+`), including cases where 'H' is followed by zero digits. Ensure that the abbreviation does not exceed three tokens.
3. **Analyze Surrounding Text**: For each identified abbreviation, examine the surrounding text to determine the corresponding full chemical name based on the specified positional relationships (e.g., `=` , `:` , `()`).
4. **Verify Accuracy**: Confirm the correct correspondence between each abbreviation and its full name to prevent errors.
5. **Organize Results**: Compile a clear and accurate list of all extracted abbreviations along with their full names, including their location in the document and the type of relationship identified.

Output Format:

Please provide the output in the following format:
Abbreviation:
Full Name:
Location:
Relationship Type:

(Repeat for each abbreviation-full name pair.)

Constraints:

- **Accuracy**:
  - **Matching Degree**: Extraction must be precise, ensuring that each abbreviation matches its full name correctly and without errors.
  - **Regular Expression Matching**: Only extract abbreviations that strictly match the provided regular expression patterns (`H\d*L\d*` , `L\d+` , `L\d+H\d+`). Do not extract abbreviations that do not conform to these patterns.
  - **Element Restrictions**: Abbreviations and full names will not contain elements representing metals from the periodic table.
- **Rigor**:
  - **No Fabrication**: Avoid extracting instances that do not exist in the original text. Do not fabricate examples that are not present in the original text.
  - **Faithfulness**: Ensure that all extracted information strictly adheres to the original document's content.
- **Efficiency**:
  - **First Occurrence Only**: Only find the first occurrence of each abbreviation-full name pair from the original text. Do not extract the same pair repeatedly.
- **Independence**:
  - **Isolated Examples**: Treat each example as independent and unrelated to others. For instance, 'H2L' may have different full names in different cases.
- **Clarity**:
  - **Well-organized Output**: The output should be well-organized and easy to understand, facilitating user comprehension and further usage.

Please process the following document:

And provide the output in the specified format.

Abbreviation:
Full Name:
Location:
Relationship Type:

(Repeat for each abbreviation-full name pair.)

Constraints:

- Do not include any additional text or explanations outside the specified output format.
- Ensure that the output is clear and follows the format exactly.

Figure S11: Prompt of chemical abbreviation resolution agent

- **Role:** You are a professional intelligent text analysis assistant with rich knowledge of organic chemistry. Your main task is to extract paragraphs focused on material synthesis from academic texts and suggest innovative modifications to existing synthesis methods when necessary.
- **Text Analysis**: Identify and extract parts related to synthesis from text, ensuring that the data is complete, quantitative, and accurate.
- **Chemical and Quantitative Knowledge**: Analyze chemical reaction details, including chemical formulas, reagent quantities, and experimental conditions, to extract accurate data.
- **Creative Adjustments**: Propose feasible modifications to synthesis processes, such as replacing reagents or optimizing conditions, while maintaining scientific integrity.

Steps

1. **Identification**:
   - Locate and extract all paragraphs related to synthesis from the document.
   - Capture details such as reagents, their quantities, chemical formulas, temperatures, durations, and any specific equipment used.
2. **Quantitative Analysis**:
   - Ensure all extracted data is quantitative, precise, and retains the scientific accuracy of the original source.
3. **Creative Synthesis Strategy**:
   - Evaluate the existing synthesis method.
   - If a modification is needed, calculate new reagent quantities based on chemical principles and propose optimized experimental conditions.
   - Replace or adjust reagents logically based on the context to ensure compatibility with different compounds, if required.
4. **Output Preparation**:
   - Compile the extracted and adjusted information clearly, suitable for direct application by researchers.

Output Format

- Present a structured and clear documentation of both the extracted synthesis data and any adjusted methods.
- Include detailed procedural steps, recalculated reagent quantities, and any modified experimental conditions.

```
**Examples**:
  - 3-shots
```

Notes

- Information must be scientifically accurate and quantitatively precise.
- Ensure synthesized information is arranged sequentially, e.g., complex (1), complex (2), etc.
- Creative adjustments must be feasible and align with contextual, chemical and experimental standards.
- Aim for comprehensive output empowering researchers to apply findings effectively.
- Exclude analysis descriptions, e.g., NMR, FT-IR data, and melting points. Watch for keywords like "Anal. Calc.", "IR (KBr pellet)," etc.
- Retain synthesis details, remove irrelevant data such as citations, chapter numbers, image and table references, or analysis information.
- If no synthesis paragraph is found, politely explain why.

Figure S12: Prompt of fine-tuning model

- **Example of 3-shots**

-**Input**:
IGUHAU
Chemical_Name: catena-((μ2-Glutarato)-bis(μ2-2-aminopyrimidine)-di-silver(i) monohydrate)
Number: 714484
Synonyms: N/A
Synthesis of complex [Ag2(L1)2(L3)]·H2O (2) A mixture of AgNO3 (170mg, 1mmol), 2-aminopyrimidine (94mg, 1mmol) and glutarate (132mg, 1mmol) were stirred in CH3OH–H2O mixed solvent (10mL, v/v: 1/1). Then aqueous NH3 solution (25%) was dropped into the mixture to give a clear solution. The resultant solution was allowed slowly to evaporate in darkness at room temperature for several days to give colorless crystals of 2 were obtained in 62% yield. They were washed with a small volume of cold CH3OH and diethyl ether.
Abbreviation: N/A
Full Name: N/A
-**Output**:
IGUHAU
Chemical_Name: catena-((μ2-Glutarato)-bis(μ2-2-aminopyrimidine)-di-silver(i) monohydrate)
Number: 714484
Synonyms: N/A
[La(H2O)2L1(NO3)3]∞ (2)

```
| Metal Source | Organic Linkers Source       | Modulator Source | Solvent Source | Quantity of Metal
| Quantity of Organic Linkers   | Quantity of Modulator | Quantity of Solvent | pH   | Synthesis
Temperature | Synthesis Time | Crystal Morphology | Yield | Equipment |
| ------------ | ---------------------------- | ----------------- | --------------- | ------------------
| ------------------------------ | ---------------------- | -------------------- | ---- | --------------
------- | -------------- | ------------------- | ----- | ---------- |
| AgNO3        | 2-aminopyrimidine; glutarate | aqueous NH3      | CH3OH–H2O      | 170 mg, 1 mmol
| 94 mg, 1 mmol; 132 mg, 1 mmol | 25%                    | 10 mL, v/v: 1/1     | N/A  | room
temperature      | several days   | colorless crystals | 62%   | N/A       |
```

Figure S13: Shot 3-1 of structured conversion agent (This case is located at 3-shots in Figure S8)

-**Input**:
SAZTIX01
Chemical_Name: catena-(bis(μ2-4,4'-Bipyridine)-(μ2-sulfato)-di-copper(i) hexahydrate)
Number: 805893
Synonyms: N/A
Cu2(4,4′-bpy)2SO4·6(H2O), (1) 0.1153g (0.46mmol) of copper sulfate, 0.1030g (0.66mmol) of 4,4′-bpy and 0.0787g (0.59mmol) of l-asp were added into 16mL of distilled water under magnetic stirring. The mixture was transferred into a 40mL Teflon-lined autoclave and heated at 120°C in an oven for 3 days. After natural cooling, the sample was filtered and washed with 3×50mL distilled water. Greenish yellow crystals of (1) were obtained as a pure phase and air-dried at ambient conditions (0.08g, yield ca. 54% based on CuSO4).
Full Name: N/A
-**Output**:
SAZTIX01
Chemical_Name: catena-(bis(μ2-4,4'-Bipyridine)-(μ2-sulfato)-di-copper(i) hexahydrate)
Number: 805893
Synonyms: N/A
Cu2(4,4′-bpy)2SO4·6(H2O), (1)

```
| Metal Source   | Organic Linkers Source | Modulator Source | Solvent Source | Quantity of Metal
| Quantity of Organic Linkers               | Quantity of Modulator | Quantity of Solvent | pH   |
Synthesis Temperature | Synthesis Time | Crystal Morphology       | Yield
| Equipment             |
| ------------- | ---------------------- | ---------------- | -------------- | --------------------
| ----------------------------------------- | ---------------------- | -------------------- | ---- | -
-------------------- | -------------- | ------------------------ | ----------------------------------
-- | ---------------------- |
| copper sulfate | 4,4′-bpy; l-asp        | N/A              | water          | 0.1153 g (0.46 mmol)
| 0.1030 g (0.66 mmol); 0.0787 g (0.59 mmol) | N/A                  | 16 mL              | N/A  |
120 °C                | 3 days         | Greenish yellow crystals | 0.08 g, yield ca. 54% based on
CuSO4 | Teflon-lined autoclave |
```

Figure S14: Shot 3-2 of structured conversion agent (This case is located at 3-shots in Figure S8)

-**Input**:
TUSVUZ
Chemical_Name: catena-[bis(μ3-Glutarato)-(μ2-aqua)-tetrapyridyl-di-cobalt(ii) pyridine solvate]
Number: 714524
Synonyms: N/A
Synthesis of complex [Ag2(L1)2(L3)]·H2O (2) A mixture of AgNO3 (170mg, 1mmol), 2-aminopyrimidine (94mg, 1mmol) and glutarate (132mg, 1mmol) were stirred in CH3OH–H2O mixed solvent (10mL, v/v: 1/1). Then aqueous NH3 solution (25%) was dropped into the mixture to give a clear solution. The resultant solution was allowed slowly to evaporate in darkness at room temperature for several days to give colorless crystals of 2 were obtained in 62% yield. They were washed with a small volume of cold CH3OH and diethyl ether.
Abbreviation: N/A
Full Name: N/A
-**Output**:
TUSVUZ
Chemical_Name: catena-[bis(μ3-Glutarato)-(μ2-aqua)-tetrapyridyl-di-cobalt(ii) pyridine solvate]
Number: 714524
Synonyms: N/A
{[Co2(μ-OH2)(μ-glutarato)2(pyridine)4]·pyridine} n (2)

```
| Metal Source                    | Organic Linkers Source | Modulator Source | Solvent Source
| Quantity of Metal  | Quantity of Organic Linkers | Quantity of Modulator | Quantity of Solvent | pH
| Synthesis Temperature | Synthesis Time | Crystal Morphology | Yield  | Equipment |
| ------------------------------ | ---------------------- | ----------------- | ---------------------
------------ | ------------------ | --------------------------- | --------------------- | -----------
-------- | ---- | ---------------------- | -------------- | ------------------- | ------ | --------- |
| cobalt(II) acetate tetrahydrate | glutaric acid          | N/A               | pyridine;
dimethylformamide (DMF) | 0.124 g (0.5 mmol) | 0.066 g (0.5 mmol)          | N/A                   |
10 cm3; 10 cm3      | N/A  | room temperature      | several days   | pink crystals      | 78.70% |
N/A       |
```

Figure S15: Shot 3-3 of structured conversion agent (This case is located at 3-shots in Figure S8)

Example of 1-shot

- **Input**:

BECSIM
Chemical_Name: catena-((μ2-1,4-bis((2-Pyridinio)methoxy)benzene-N,N'-diacetato-O,O')-diaqua-tris(nitrato-O,O')-lanthanum(iii))
Number: 207467
Synonyms: N/A
Synthesis of [La(H2O)2L1(NO3)3]∞ (2) An aqueous solution (4 ml) containing La(NO3)3·6H2O (0.433 g, 1 mmol) was mixed with an aqueous solution (6 ml) of L1(0.08 g, 0.2 mmol), and the mixture was stirred at 70°C for about 10 min and then filtered. The filtrate was allowed to stand at room temperature. Several days later, colorless block-like crystals of 2 were obtained. Yield: 37%.
Abbreviation: L1
Full Name: 1,4-bis(2-picolyloxyl)benzene-N,N′-diacetate

- **Output**:

BECSIM
Chemical_Name: catena-((μ2-1,4-bis((2-Pyridinio)methoxy)benzene-N,N'-diacetato-O,O')-diaqua-tris(nitrato-O,O')-lanthanum(iii))
Number: 207467
Synonyms: N/A
[La(H2O)2L1(NO3)3]∞ (2)

Metal Source	Organic Linkers Source	Modulator Source	Solvent Source	Quantity of Metal	Quantity of Organic Linkers	Quantity of Modulator	Quantity of Solvent	pH	Synthesis Temperature	Synthesis Time	Crystal Morphology	Yield	Equipment
La(NO3)3·6H2O	1,4-bis(2-picolyloxyl)benzene-N,N′-diacetate	N/A	aqueous solution	0.433 g, 1 mmol	0.08 g, 0.2 mmol	N/A	10mL	N/A	room temperature	Several days	colorless block-like crystals	37%	N/A

Figure S16: Shot 1-1 of structured conversion agent (This case is located at 1-shot in Figure S8)

- **5-shots of Contents Table**:

1. **Text**: Table 1 Crystal data and structure refinement details for Ba2M(CHOO)6(H2O)4, (M=Co, Ni, Cu, Zn) M=Co (1) M=Ni (2) M=Cu (3) M=Zn (4) Empirical formula C6 H14 Ba2 Co O16 C6 H14 Ba2 Ni O16 C6 H14 Ba2 Cu O16 C6 H14 Ba2 O16 Zn Formula weight 675.78 675.56 680.39 682.22 a (Å) 8.9191(7) 8.8809(7) 8.8166(9) 8.9357(8) b (Å) 7.1394(6) 7.1281(5) 7.1279(7) 7.1394(6) c (Å) 6.9062(6) 6.8970(5) 6.9281(7) 6.9018(6) α (°) 99.4770(10) 99.3400(10) 98.000(2) 99.4100(10) β (°) 108.8490(10) 108.8650(10) 108.9400(10) 108.7820(10) γ (°) 82.4050(10) 82.4450(10) 82.546(2) 82.4530(10) V (Å3) 409.03(6) 406.21(5) 406.14(7) 409.79(6) D calc (gcm−3) 2.743 2.762 2.782 2.765 μ (mm−1) 5.845 6.023 6.175 6.286 F(000) 317 318 319 320 Crystal shape, color Prisms, pale rose Plates, light green Plates, light blue Prisms, colorless θ range (°) 2.42–27.72 2.43–27.94 2.45–27.75 2.42–28.05 Index ranges −11<=h<=11, −8<=k<=9, −8<=l<=9 −11<=h<=11, −9<=k<=9, −8<=l<=8 −11<=h<=11, −9<=k<=9, −8<=l<=9 −11<=h<=11, −9<=k<=9, −9<=l<=8 N tot, N uniq(R int), N [I>2σ(I)] 3411, 1749(0.012), 1688 3389, 1740(0.011), 1710 3346, 1717(0.0182), 1676 3423, 1769(0.0186), 1707 Goodness-of-fit S 0.931 0.987 1.071 0.932 R indices [I>2σ(I)] R1=0.0180, wR2=0.0497 R1=0.0179, wR2=0.0506 R1=0.0223, wR2=0.0570 R1=0.0201, wR2=0.0525 R indices (all data) R1=0.0190, wR2=0.0504 R1=0.0183, wR2=0.0509 R1=0.0239, wR2=0.0595 R1=0.0211, wR2=0.0534 ρ max/ρ min (eÅ−3) 0.409 and −0.891 0.425 and −0.890 0.977 and −1.165 0.919 and −0.821 Details in common: T (K): 293(2), λ (Å): 0.71073, crystal system: triclinic, space group: P−1, Z: 1, parameters refined: 268, absorption correction: Multi_scan.

   - **Extraction**:


```
[
  {
      "Complex": "M=Co (1)",
      "Empirical formula": "C6H14Ba2CoO16",
      "Formula weight": "675.78",
      "a (Å)": "8.9191(7)",
      "b (Å)": "7.1394(6)",
      "c (Å)": "6.9062(6)",
      "alpha (°)": "99.4770(10)",
      "beta (°)": "108.8490(10)",
      "gamma (°)": "82.4050(10)",
      "Crystal shape, color": "Prisms, pale rose"
  },
  {
      "Complex": "M=Ni (2)",
      "Empirical formula": "C6H14Ba2NiO16",
      "Formula weight": "675.56",
      "a (Å)": "8.8809(7)",
      "b (Å)": "7.1281(5)",
      "c (Å)": "6.8970(5)",
      "alpha (°)": "99.3400(10)",
      "beta (°)": "108.8650(10)",
      "gamma (°)": "82.4450(10)",
      "Crystal shape, color": "Plates, light green"
  },
  {
      "Complex": "M=Cu (3)",
      "Empirical formula": "C6H14Ba2CuO16",
      "Formula weight": "680.39",
      "a (Å)": "8.8166(9)",
      "b (Å)": "7.1279(7)",
      "c (Å)": "6.9281(7)",
      "alpha (°)": "98.000(2)",
      "beta (°)": "108.9400(10)",
      "gamma (°)": "82.546(2)",
      "Crystal shape, color": "Plates, light blue"
  },
  {
      "Complex": "M=Zn (4)",
      "Empirical formula": "C6H14Ba2O16Zn",
      "Formula weight": "682.22",
      "a (Å)": "8.9357(8)",
      "b (Å)": "7.1394(6)",
      "c (Å)": "6.9018(6)",
```

```
        "alpha (°)": "99.4100(10)",
        "beta (°)": "108.7820(10)",
        "gamma (°)": "82.4530(10)",
        "Crystal shape, color": "Prisms, colorless"
    }
]
```

Figure S17: Shot 5-1 of table data parsing agent (This case is located at 5-shots in Figure S9)

2. **Text**: Table 1 Crystallographic data for 1, 2, 3 and 4 Complexes 1 2 3 4 Empirical formula C80H64N8O22Mn2 C32H27N4O7.5Mn C18H16N2O6Mn C16H18N2O8Mn Formula weight 1559.27 642.51 411.27 421.26 Crystal system orthorhombic triclinc triclinic triclinic Colour light yellow light yellow dark yellow light yellow Space group Pbca P 1 ¯ P 1 ¯ P 1 ¯ a (Å) 12.9669(5) 11.4770(5) 7.149(1) 8.0850(7) b (Å) 22.2619(9) 13.0536(6) 10.171(2) 8.1420(7) c (Å) 25.671(1) 21.2630(6) 11.936(2) 13.680(1) α (°) 104.880(1) 88.657(4) 81.040(1) β (°) 90.770(1) 79.007(5) 87.130(2) γ (°) 107.790(1) 89.083(6) 77.910(2) V (Å3) 7410.3(5) 2916.7(2) 851.7(3) 869.7(1) Z 4 4 2 2 D calc. (gcm−3) 1.433 1.463 1.604 1.609 Crystal size (mm) 0.24×0.48×0.56 0.05×0.18×0.26 0. 09×0.09×0.10 0.16×0.15×0.11 θ Range (°) 1.59–25.00 1.00–24.00 1.74–28.30 1.51–25.00 T (K) 296(2) μ (mm−1) 0.425 0.511 0.815 0.808 λ (Mo Kα) (Å) 0.7107 Reflections collected/unique 68221/6535 25162/9140 5895/3716 5074/2774 Data/parameters 6535/505 9140/802 3716/256 2774/268 Goodness-of-fiton F 2 1.034 1.041 0.934 1.088 R [I >2σ(I)] R 1 =0.0615 R 1 =0.0923 R 1 =0.0530, R 1 =0.0607 wR 2 =0.2040 wR 2 =0.1950 wR 2 =0.0762 wR 2 =0.1707 R(all data) R 1 =0.0710 R 1 =0.1458 R 1 =0.1148 R 1 =0.0747 wR 2 =0.2133 wR 2 =0.2359 wR 2 =0.0873 wR 2 =0.2060 Δρ min and Δρ max (e Å−3) −0.537, 1.359 −0.909, 0.961 −0.715, 0.434 −0.566, 0.718

- **Extraction**:


```
[
    {
        "Complexes": "1",
        "Empirical formula": "C80H64N8O22Mn2",
        "Formula weight": "1559.27",
        "Crystal system": "orthorhombic",
        "Colour": "light yellow",
        "Space group": "Pbca",
        "a (Å)": "12.9669(5)",
        "b (Å)": "22.2619(9)",
        "c (Å)": "25.671(1)",
        "alpha (°)": "NA",
        "beta (°)": "NA",
        "gamma (°)": "NA"
    },
    {
        "Complexes": "2",
        "Empirical formula": "C32H27N4O7.5Mn",
        "Formula weight": "642.51",
        "Crystal system": "triclinic",
        "Colour": "light yellow",
        "Space group": "P1¯",
        "a (Å)": "11.4770(5)",
        "b (Å)": "13.0536(6)",
        "c (Å)": "21.2630(6)",
        "alpha (°)": "104.880(1)",
        "beta (°)": "90.770(1)",
        "gamma (°)": "107.790(1)"
    },
    {
        "Complexes": "3",
        "Empirical formula": "C18H16N2O6Mn",
        "Formula weight": "411.27",
        "Crystal system": "triclinic",
        "Colour": "dark yellow",
        "Space group": "P1¯",
        "a (Å)": "7.149(1)",
        "b (Å)": "10.171(2)",
        "c (Å)": "11.936(2)",
        "alpha (°)": "88.657(4)",
        "beta (°)": "79.007(5)",
        "gamma (°)": "89.083(6)"
    },
    {
        "Complexes": "4",
        "Empirical formula": "C16H18N2O8Mn",
        "Formula weight": "421.26",
        "Crystal system": "triclinic",
        "Colour": "light yellow",
        "Space group": "P1¯",
```

```
        "a (Å)": "8.0850(7)",
        "b (Å)": "8.1420(7)",
        "c (Å)": "13.680(1)",
        "alpha (°)": "81.040(1)",
        "beta (°)": "87.130(2)",
        "gamma (°)": "77.910(2)"
    }
]
```


Figure S18: Shot 5-2 of table data parsing agent (This case is located at 5-shots in Figure S9)

3. **Text**: Table 1 Crystallographic data for compounds 1–3 a 1 2 3 Empirical formula C6H15AgClN3O4 C6H15AgClN3O4 C6H15AgBF4N3 Formula weight 336.53 336.53 323.89 Crystal system monoclinic monoclinic monoclinic a (Å) 8.4778(5) 7.3136(8) 8.4228(4) b (Å) 7.6947(3) 8.0334(13) 7.6234(4) c (Å) 16.6640(12) 8.9245(12) 16.4719(8) α (°) 90 90 90 β (°) 96.803(2) 101.289(8) 97.792(3) γ (°) 90 90 90 Space group P21/c P21/c P21/c V (Å3) 1079.41(11) 514.20(12) 1047.90(9) Z value 4 2 4 ρ calc (g/cm3) 2.071 2.174 2.053 μ (cm−1) 2.114 2.219 1.950 T (K) 150(2) 150(2) 150(2) Number of observations 2121 1559 2061 Residuals: R;R w 0.0295; 0.0578 0.0590; 0.1041 0.0308; 0.0500 a R1=∑‖F o |−|F c ‖ ∑|F o |. wR2= ∑[w(F o 2−F c 2)2] ∑[w(F o 2)2] 1/2 .

- **Extraction**:


```
[
    {
        "Complexes": "1",
        "Empirical formula": "C6H15AgClN3O4",
        "Formula weight": "336.53",
        "Crystal system": "monoclinic",
        "a (Å)": "8.4778(5)",
        "b (Å)": "7.6947(3)",
        "c (Å)": "16.6640(12)",
        "alpha (°)": "90",
        "beta (°)": "96.803(2)",
        "gamma (°)": "90",
        "Space group": "P21/c"
    },
  {
        "Complexes": "2",
        "Empirical formula": "C6H15AgClN3O4",
        "Formula weight": "336.53",
        "Crystal system": "monoclinic",
        "a (Å)": "7.3136(8)",
        "b (Å)": "8.0334(13)",
        "c (Å)": "8.9245(12)",
        "alpha (°)": "90",
        "beta (°)": "101.289(8)",
        "gamma (°)": "90",
        "Space group": "P21/c"
    },
    {
        "Complexes": "3",
        "Empirical formula": "C6H15AgBF4N3",
        "Formula weight": "323.89",
        "Crystal system": "monoclinic",
        "a (Å)": "8.4228(4)",
        "b (Å)": "7.6234(4)",
        "c (Å)": "16.4719(8)",
        "alpha (°)": "90",
        "beta (°)": "97.792(3)",
        "gamma (°)": "90",
        "Space group": "P21/c"
    }
]
```


Figure S19: Shot 5-3 of table data parsing agent (This case is located at 5-shots in Figure S9)

4. **Text**: Table 4 Crystallographic data of the ligand and complexes 1·2CH3CN and 2 dpim 1·2CH3CN 2 Formula C16H15N2P Cu2C54H54N9P3Cl2O8 Ag2C32H30N6P2O6 Formula weight 266.28 1248.00 872.31 Crystal dimension (mm) 0.20×0.20×0.50 0.60×0.15×0.10 0.30×0.10×0.50 Crystal system monoclinic monoclinic monoclinic Space group P21/c (no. 14) P21/c (no. 14) P2/c (no. 13) Unit cell dimensions a (Å) 13.283(2) 14.215(2) 13.814(4) b (Å) 9.917(2) 24.872(6) 9.338(2) c (Å) 10.864(3) 16.093(3) 14.974(4) β (°) 93.59(2) 91.610(9) 116.21(2) U (Å3) 1428.3(4) 5687(1) 1733.0(7) Z 4 4 2 μ(Mo Kα) (cm−1) 1.80 9.87 12.71 T (°C) 23.0 −120.0 23 Reflections (total) 3651 8467 4382 Reflections (unique) 3476 4224 R int 0.017 0.053 R; R w 0.038; 0.059 0.039; 0.054 0.079; 0.123

- **Extraction**:


```
[
    {
        "Complexes": "dpim",
        "Formula": "C16H15N2P",
        "Formula weight": "266.28",
        "Crystal system": "monoclinic",
        "Space group": "P21/c (no. 14)",
        "a (Å)": "13.283(2)",
        "b (Å)": "9.917(2)",
        "c (Å)": "10.864(3)",
        "beta (°)": "93.59(2)"
    },
    {
        "Complexes": "1·2CH3CN",
        "Formula": "Cu2C54H54N9P3Cl2O8",
        "Formula weight": "1248.00",
        "Crystal system": "monoclinic",
        "Space group": "P21/c (no. 14)",
        "a (Å)": "14.215(2)",
        "b (Å)": "24.872(6)",
        "c (Å)": "16.093(3)",
        "beta (°)": "91.610(9)"
    },
    {
        "Complexes": "2",
        "Formula": "Ag2C32H30N6P2O6",
        "Formula weight": "872.31",
        "Crystal system": "monoclinic",
        "Space group": "P2/c (no. 13)",
        "a (Å)": "13.814(4)",
        "b (Å)": "9.338(2)",
        "c (Å)": "14.974(4)",
        "beta (°)": "116.21(2)"
    }
]
```


Figure S20: Shot 5-4 of table data parsing agent (This case is located at 5-shots in Figure S9)

5. **Text**:

Table 1S. Crystallographic data for 1- 6
Compound 1 2 3 4 5 6
Empirical formula C H N O Sr C H N O Zn C H CuN O C H N O Pb C H ClEuN O C H ClN O Tb
16 18 10 7 16 20 10 8 16 18 10 7 16 12 10 4 16 16 10 6 16 16 10 6
Formula mass 550.02 545.81 525.95 615.55 631.80 638.77
Crystal system monoclinic triclinic monoclinic monoclinic monoclinic monoclinic
Space group P2 /c P ī P2 P2 /c C2/c C2/c
1 1 1
a (Å) 8.0445(16) 7.1354(14) 7.8099(16) 12.768(3) 23.514(5) 23.432(5)
b (Å) 21.141(4) 9.3422(19)) 14.717(3) 8.1286(16) 11.223(2) 11.161(2)
c (Å) 13.565(3 10.662(2) 9.3942(19) 20.797(4) 8.8121(18) 8.8797(18)
α (°) 64.51(3)
 (°) 92.60(3) 72.84(3) 108.53(3) 97.81(3) 107.66(3) 107.95(3)
 (°) 79.91(3)
V (Å3) 2304.6(8) 612.0(2) 1023.8(4) 2138.4(7) 2215.9(8) 2209.2(8)
Z 4 1 2 4 4 4
T/K 291(2) 291(2) 291(2) 291(2) 291(2) 291(2)
D
calcd
(g.cm-3) 1.585 1.481 1.706 1.912 1.894 1.920
 (mm-1) 2.398 1.064 1.133 7.935 3.008 3.379
Reflections 8686 10777 10869
21190 6435 16535
collected
Unique 3598 (0.0918) 2541 (0.0339) 2527 (0.0415)
4691(0.1073) 2794 (0.0329) 3760 (0.1023)
Reflections(R)
int
No. Observations (I 2875 2138 2193
3341 2347 2270
>2.00 (I))
308 280 164 153
No. Variables 307 165
R[a] , wR[b] 0.0726, 0.1642 0.0496, 0.1339 0.0878, 0.1703 0.0623, 0.1036 0.0738, 0.2119 0.0440, 0.1226

..

- **Extraction**:


```
[
  {
      "Compound": "1",
      "Empirical formula": "C16H18N10O7Sr",
      "Formula mass": "550.02",
      "Crystal system": "monoclinic",
      "Space group": "P2/c",
      "a (Å)": "8.0445(16)",
      "b (Å)": "21.141(4)",
      "c (Å)": "13.565(3)",
      "alpha (°)": "N/A",
      "beta (°)": "92.60(3)",
      "gamma (°)": "N/A"
  },
  {
      "Compound": "2",
      "Empirical formula": "C16H20N10O8Zn",
      "Formula mass": "545.81",
      "Crystal system": "triclinic",
      "Space group": "Pī",
      "a (Å)": "7.1354(14)",
      "b (Å)": "9.3422(19)",
      "c (Å)": "10.662(2)",
      "alpha (°)": "64.51(3)",
```

```
        "beta (°)": "72.84(3)",
        "gamma (°)": "79.91(3)"
    },
    {
        "Compound": "3",
        "Empirical formula": "C16H18CuN10O7",
        "Formula mass": "525.95",
        "Crystal system": "monoclinic",
        "Space group": "P2",
        "a (Å)": "7.8099(16)",
        "b (Å)": "14.717(3)",
        "c (Å)": "9.3942(19)",
        "alpha (°)": "N/A",
        "beta (°)": "108.53(3)",
        "gamma (°)": "N/A"
    },
  {
        "Compound": "4",
        "Empirical formula": "C16H12N10O4Pb",
        "Formula mass": "615.55",
        "Crystal system": "monoclinic",
        "Space group": "P2/c",
        "a (Å)": "12.768(3)",
        "b (Å)": "8.1286(16)",
        "c (Å)": "20.797(4)",
        "alpha (°)": "N/A",
        "beta (°)": "97.81(3)",
        "gamma (°)": "N/A"
    },
    {
        "Compound": "5",
        "Empirical formula": "C16H16ClEuN10O6",
        "Formula mass": "631.80",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "23.514(5)",
        "b (Å)": "11.223(2)",
        "c (Å)": "8.8121(18)",
        "alpha (°)": "N/A",
        "beta (°)": "107.66(3)",
        "gamma (°)": "N/A"
    },
    {
        "Compound": "6",
        "Empirical formula": "C16H16ClN10O6Tb",
        "Formula mass": "638.77",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "23.432(5)",
        "b (Å)": "11.161(2)",
        "c (Å)": "8.8797(18)",
        "alpha (°)": "N/A",
        "beta (°)": "107.95(3)",
        "gamma (°)": "N/A"
    }
]
```


Figure S21: Shot 5-5 of table data parsing agent (This case is located at 5-shots in Figure S9)

- **3-shots Contents Crystal data**:

1. **Text**: Crystal data for compound 1: Orthorhombic, Fdd2, a=20.0763(17)Å, b=54.982(5)Å, c=9.0363(8)Å, α =90°, β =90°, γ =90°, V=9974.6(15)Å3, Z=16, Dc=2.128, F(000)=6160.0, GOF=1.035, R1(I>2σ(I))=0.0335, and wR2 (all data)=0.0800. 2: Orthorhombic, Fdd2, a=20.0763(17)Å, b=54.982(5)Å, c=9.0363(8)Å, α =90°, β =90°, γ =90°, V=9974.6(15)Å3, Z=16, Dc=2.140, F(000)=6192.0, GOF=1.011, R1(I>2σ(I))=0.0391, and wR2 (all data)=0.0873. 3: Orthorhombic, Fdd2, a=20.106(3)Å, b=55.082(9)Å, c=9.0499(14)Å, α =90°, β =90°, γ =90°, V=10023(3)Å3, Z=16, Dc=2.124, F(000)=6176.0, GOF=1.045, R1(I>2σ(I))=0.0321, and wR2 (all data)=0.0691.

- **Extraction**:


```
[
    {
  "Compound": "1",
  "Crystal system": "Orthorhombic",
  "Space group": "Fdd2",
  "a (Å)": "20.0763(17)",
  "b (Å)": "54.982(5)",
  "c (Å)": "9.0363(8)",
  "alpha (°)": "90",
  "beta (°)": "90",
  "gamma (°)": "90"
    },
    {
  "Compound": "2",
  "Crystal system": "Orthorhombic",
  "Space group": "Fdd2",
  "a (Å)": "20.0763(17)",
  "b (Å)": "54.982(5)",
  "c (Å)": "9.0363(8)",
  "alpha (°)": "90",
  "beta (°)": "90",
  "gamma (°)": "90"
    },
    {
  "Compound": "3",
  "Crystal system": "Orthorhombic",
  "Space group": "Fdd2",
  "a (Å)": "20.106(3)",
  "b (Å)": "55.082(9)",
  "c (Å)": "9.0499(14)",
  "alpha (°)": "90",
  "beta (°)": "90",
  "gamma (°)": "90"
    }
]
```


Figure S22: Shot 3-1 of table data parsing agent (This case is located at 3-shots in Figure S9)

2. **Text**: Single crystals of 1 and 2 were mounted on a Bruker SMART APEX CCD-based diffractometer and a Bruker P4 diffractometer equipped with the SMART CCD system, respectively. X-ray data for 1 and 2 were collected at 173(2)K and using MoKα radiation (λ =0.71073Å, graphite monochromator). The raw data were processed to give structure factors using the Bruker SAINT program and corrected for Lorentz and polarization effects. No absorption corrections were applied. The crystal structures were solved by direct methods, and refined by full-matrix least-squares refinement using the SHELXL97 computer program. All non-hydrogen atoms were refined anisotropically. All hydrogen atoms were positioned geometrically and refined using a riding model. Crystallographic data for C8H7AgF3N3O3 (1): M =358.04g/mol, pale yellow needle, 0.05×0.12×0.50mm, T =173(2)K, triclinic, space group P 1 ¯ (No. 2), a =7.3219(19)Å, b =8.609(2)Å, c =9.734(3)Å, α =79.710(5)°, β =81.029(5)°, γ =65.453(4)°, V =546.9(3)Å3, Z =2, D c =2.174g/cm3, μ =1.890mm−1, F(000)=348, GOF=1.185. Of the 3963 reflections, 2649 were unique (R int =0.0480), 1991 observed [I >2σ(I)], which refined to R1=0.0560, wR2=0.1140, and for all data R1=0.0997, wR2=0.1902. Crystallographic data for C15H18Ag2F6N6O9S2 (2): M =820.21g/mol, off-white, 0.05×0.20×0.20mm, T =173(2)K, triclinic, space group P 1 ¯ (No. 2), a =7.8433(7)Å, b =12.2795(11)Å, c =13.5977(12)Å, α =101.596(2)°, β =93.864(2)°, γ =93.200(2)°, V =1276.7(2)Å3, Z =2, D c =2.134g/cm3, μ =1.799mm−1, F(000)=804, GOF=1.135. Of the 11,566 reflections, 5863 were unique (R int =0.0541), 4781 observed [I >2σ(I)], which refined to R1=0.0439, wR2=0.1043, and for all data R1=0.0587, wR2=0.1194.

- **Extraction**:


```
[
    {
        "Compound": "1",
        "Empirical formula": "C8H7AgF3N3O3",
        "Formula weight (M)": "358.04",
        "Crystal system": "triclinic",
        "Space group": "P1¯ (No. 2)",
        "a (Å)": "7.3219(19)",
        "b (Å)": "8.609(2)",
        "c (Å)": "9.734(3)",
        "alpha (α)": "79.710(5)",
        "beta (β)": "81.029(5)",
        "gamma (γ)": "65.453(4)",
        "Colour": "pale yellow needle"
    },
    {
        "Compound": "2",
        "Empirical formula": "C15H18Ag2F6N6O9S2",
        "Formula weight (M)": "820.21",
        "Crystal system": "triclinic",
        "Space group": "P1¯ (No. 2)",
        "a (Å)": "7.8433(7)",
        "b (Å)": "12.2795(11)",
        "c (Å)": "13.5977(12)",
        "alpha (α)": "101.596(2)",
        "beta (β)": "93.864(2)",
        "gamma (γ)": "93.200(2)",
        "Colour": "off-white"
    }
]
```


Figure S23: Shot 3-2 of table data parsing agent (This case is located at 3-shots in Figure S9)

3. **Text**: [11] IR (KBr, cm−1): ν =3459m, 3245w, 1639s, 1427m, 1327m, 1121s, 928w, 799m, 613m, 507m. Anal. Calcd: C, 7.76; H, 1.94; N, 4.53. Found: C, 7.81; H, 1.98; N, 4.62; yield: 80% based on metal oxide. Crystal data for 1: FW, 309.05; Monoclinic; space group, P2(1)/ n; cell dimensions: a =6.5341(13)Å, b =8.5266(17)Å, c =13.745(3)Å, β =92.87(3)°, V =764.8(3)Å3, Z =4, μ =7.934mm−1, T =293(2)K. Reflections collected: 7330; independent reflections: 1738; Final R indices [I>2σ(I)]: R a =0.0253, R w =0.0680. [12] IR (KBr, cm−1): ν =3440m, 3041w, 1635s, 1420m, 1321m, 1110s, 920w, 797m, 598m, 501m. Anal. Calcd: C, 6.68; H, 1.67; N, 3.90. Found: C, 6.70; H, 1.72; N, 3.97; yield: 74% based on ligand. Crystal data for 2: FW, 359.05; Monoclinic; space group, P2(1)/ n; cell dimensions: a =6.6294(13)Å, b =8.6727(17)Å, c =13.867(3)Å, β =93.35(3)°, V =795.9(3)Å3, Z =4, μ =5.664mm−1, T =293(2)K. Reflections collected: 7474; independent reflections: 1823; Final R indices [I>2σ(I)]: R a =0.0550, R w =0.1516. [13] IR (KBr, cm−1): ν =3445m, 3040w, 1672s, 1608s, 1465w, 1320m, 1146s, 981w, 789m, 653m, 611m, 494m. Anal. Calcd: C, 6.48; H, 1.62; N, 3.78. Found: C, 6.56; H, 1.65; N, 3.82; yield: 84% based on ligand. data for 3: FW, 370.49; Monoclinic; space group, P2(1)/ n; cell dimensions: a =6.6189(13)Å, b =8.6598(17)Å, c =13.853(3)Å, β =93.30(3)°, V =792.7(3)Å3, Z =4, μ =7.706mm−1, T =293(2)K. Reflections collected: 7208; independent reflections: 1819; Final R indices [I>2σ(I)]: R a =0.0319, R w =0.0835.

- **Extraction**:


```
[
  {
    "Compound": "1",
    "Formula weight (FW)": "309.05",
    "Crystal system": "Monoclinic",
    "Space group": "P2(1)/n",
    "a (Å)": "6.5341",
    "b (Å)": "8.5266(17)",
    "c (Å)": "13.745(3)",
    "beta (β)": "92.87(3)"
  },
  {
    "Compound": "2",
    "Formula weight (FW)": "359.05",
    "Crystal system": "Monoclinic",
    "Space group": "P2(1)/n",
    "a (Å)": "6.6294(13)",
    "b (Å)": "8.6727(17)",
    "c (Å)": "13.867(3)",
    "beta (β)": "93.35(3)"
  },
  {
    "Compound": "3",
    "Formula weight (FW)": "370.49",
    "Crystal system": "Monoclinic",
    "Space group": "P2(1)/n",
    "a (Å)": "6.6189(13)",
    "b (Å)": "8.6598(17)",
    "c (Å)": "13.853(3)",
    "beta (β)": "93.30(3)"
  }
]
```


Figure S24: Shot 3-3 of table data parsing agent (This case is located at 3-shots in Figure S9)

- **Examples of 5-shots:**

1. **Example 1**:


```
  - **a.json**:
    {
      "CEDDIZ": {
          "Molecule_Identifier": "CEDDIZ",
          "Number": 237610,
          "Chemical_Name": "catena-(4,4'-bipyridylium (μ2-4,4'-bipyridyl)-tetrabromo-bismuth(iii))",
          "Chemical_Name_HTML": "catena-(4,4'-bipyridylium (μ<sub>2</sub>-4,4'-bipyridyl)-tetrabromo-
bismuth(iii))",
          "Synonyms": [],
          "Crystal_System": "monoclinic",
          "Spacegroup_Symbol": "Cc",
          "a": 12.167300000000001,
          "b": 14.927800000000001,
          "c": 13.8496,
          "alpha(α)": 90.0,
          "beta(β)": 115.93400000000001,
          "gamma(γ)": 90.0,
          "Formula": "(C10 H8 Bi1 Br4 N2 1-)n,n(C10 H9 N2 1+)",
          "Heaviest_Component_Identifier": "01",
          "Heaviest_Component_Molecular_Weight": 907.7671,
          "Component_Weights": [
              907.7671,
              157.1915
          ],
          "Molecular_Weight": 1064.9586,
          "Color": "light yellow",
          "Melting_Point": "493 K",
          "R_Factor": 3.95,
          "Solvent_Names": [],
          "Publication_Details": "Citation(authors='A.Morsali', journal='Journal(Solid State
Sciences)', volume='8', year=2006, first_page='82', doi='10.1016/j.solidstatesciences.2005.10.004')"
      }
    }
  - **b.json**:"CEDDIZ": [
    {
        "Compound": "Complex",
        "Empirical formula": "C20H17Br4BiN4",
        "Formula weight": "842.00",
        "Crystal system": "monoclinic",
        "Space group": "Cc",
        "a (Å)": "12.1673 (9)",
        "b (Å)": "14.9278 (11)",
        "c (Å)": "13.8496 (11)",
        "beta (°)": "115.634 (2)"
    }
    ]
```


   - **Output**:


```
{
  "CEDDIZ":
    {
        "Compound": Complex,
        "Molecule_Identifier": CEDDIZ,
        "Number": 237610,
        "Chemical_Name": catena-(4,4'-bipyridylium (μ2-4,4'-bipyridyl)-tetrabromo-bismuth(iii)),
        "Synonyms": []
    }
}
```


Figure S25: Shot 5-1 of crystal data comparison agent (This case is located at 5-shots in Figure S10)

2. **Example 2**:


```
  - **a.json**:
    {
      "CEGWUH": {
          "Molecule_Identifier": "CEGWUH",
          "Number": 294238,
          "Chemical_Name": "catena-(bis(μ2-Thiocyanato-N,S)-bis(2,5-bis(2-pyrazinyl)-1,3,4-
oxadiazole-N)-cadmium(ii))",
          "Chemical_Name_HTML": "catena-(bis(μ<sub>2</sub>-Thiocyanato-N,S)-bis(2,5-bis(2-
pyrazinyl)-1,3,4-oxadiazole-N)-cadmium(ii))",
          "Synonyms": [],
          "Crystal_System": "triclinic",
          "Spacegroup_Symbol": "P-1",
          "a": 5.688,
          "b": 8.916,
          "c": 13.176,
          "alpha(α)": 71.365,
          "beta(β)": 82.003,
          "gamma(γ)": 88.528,
          "Formula": "(C22 H12 Cd1 N14 O2 S2)n",
          "Heaviest_Component_Identifier": "01",
          "Heaviest_Component_Molecular_Weight": 933.7992,
          "Component_Weights": [
              933.7992
          ],
          "Molecular_Weight": 933.7992,
          "Color": "yellow",
          "Melting_Point": null,
          "R_Factor": 2.3000000000000003,
          "Solvent_Names": [],
          "Publication_Details": "Citation(authors='Miao Du, Cheng-Peng Li, Jian-Hua Guo',
journal='Journal(Inorganica Chimica Acta)', volume='359', year=2006, first_page='2575',
doi='10.1016/j.ica.2006.02.039')"
      }
    }
  - **b.json**: "CEGWUH": [
    {
        "Compound": "bpzo",
        "Empirical formula": "C10H6N6O",
        "Formula weight (M)": "226.21",
        "Crystal size (mm)": "0.37×0.16×0.08",
        "Crystal system": "monoclinic",
        "Space group": "P21/n",
        "a (Å)": "3.8164(7)",
        "b (Å)": "17.175(3)",
        "c (Å)": "14.997(3)",
        "alpha (°)": "90",
        "beta (°)": "93.510(2)",
        "gamma (°)": "90",
        "V (Å3)": "981.1(3)",
        "Z": "4",
        "D calc (gcm-3)": "1.531",
        "μ (mm-1)": "0.110",
        "F(000)": "464",
        "R indices (I > 2σ(I))": "0.0315, 0.0905",
        "Goodness-of-fit on F2": "1.066"
    },
    {
        "Compound": "1",
        "Empirical formula": "C22H12CdN14O2S2",
        "Formula weight (M)": "680.98",
        "Crystal size (mm)": "0.40×0.20×0.18",
        "Crystal system": "triclinic",
        "Space group": "P 1 ¯",
        "a (Å)": "5.688(2)",
```

```
        "b (Å)": "8.916(3)",
        "c (Å)": "13.176(5)",
        "alpha (°)": "90",
        "beta (°)": "82.003(5)",
        "gamma (°)": "88.528(6)",
        "V (Å3)": "626.9(4)",
        "Z": "1",
        "D calc (gcm-3)": "1.804",
        "μ (mm-1)": "1.092",
        "F(000)": "338",
        "R indices (I > 2σ(I))": "0.0230, 0.0559",
        "Goodness-of-fit on F2": "1.074"
    },
    {
        "Compound": "2",
        "Empirical formula": "C26H26CoN16O6S2",
        "Formula weight (M)": "781.68",
        "Crystal size (mm)": "0.20×0.16×0.12",
        "Crystal system": "triclinic",
        "Space group": "P 1 ¯",
        "a (Å)": "8.514(3)",
        "b (Å)": "8.717(3)",
        "c (Å)": "12.807(4)",
        "alpha (°)": "83.241(5)",
        "beta (°)": "89.667(5)",
        "gamma (°)": "68.436(5)",
        "V (Å3)": "877.1(5)",
        "Z": "1",
        "D calc (gcm-3)": "1.480",
        "μ (mm-1)": "0.672",
        "F(000)": "401",
        "R indices (I > 2σ(I))": "0.0360, 0.0944",
        "Goodness-of-fit on F2": "1.045"
    }
    ]
```


- **Output**:


```
{
  "CEGWUH":
    {
        "Compound": 1,
        "Molecule_Identifier": CEGWUH,
        "Number": 294238,
        "Chemical_Name": catena-(bis(μ2-Thiocyanato-N,S)-bis(2,5-bis(2-pyrazinyl)-1,3,4-oxadiazole-N)-
cadmium(ii)),
        "Synonyms": []
    }
}
```


Figure S26: Shot 5-2 of crystal data comparison agent (This case is located at 5-shots in Figure S10)

3. **Example 3**:


```
  - **a.json**:
    {
      "CEGPEK": {
          "Molecule_Identifier": "CEGPEK",
          "Number": 283271,
          "Chemical_Name": "catena-(tris(μ3-Tartrato)-tris(μ2-aqua)-hexa-aqua-hexa-lithium (μ3-
tartrato)-(μ2-aqua)-diaqua-di-lithium)",
          "Chemical_Name_HTML": "catena-(tris(μ<sub>3</sub>-Tartrato)-tris(μ<sub>2</sub>-aqua)-hexa-
aqua-hexa-lithium (μ<sub>3</sub>-tartrato)-(μ<sub>2</sub>-aqua)-diaqua-di-lithium)",
          "Synonyms": [],
          "Crystal_System": "monoclinic",
          "Spacegroup_Symbol": "P21/c",
          "a": 20.8669,
          "b": 15.0074,
          "c": 11.678500000000001,
          "alpha(α)": 90.0,
          "beta(β)": 99.367,
          "gamma(γ)": 90.0,
          "Formula": "(C12 H30 Li6 O27)n,n(C4 H10 Li2 O9)",
          "Heaviest_Component_Identifier": "01",
          "Heaviest_Component_Molecular_Weight": 777.803,
          "Component_Weights": [
              777.803,
              280.9023
          ],
          "Molecular_Weight": 1058.7053,
          "Color": "colorless",
          "Melting_Point": null,
          "R_Factor": 4.97,
          "Solvent_Names": [],
          "Publication_Details": "Citation(authors='T.Gelbrich, T.L.Threlfall, S.Huth, E.Seeger',
journal='Journal(Polyhedron)', volume='25', year=2006, first_page='937',
doi='10.1016/j.poly.2005.10.021')"
      }
    }
  - **b.json**: "CEGPEK": [
    {
        "Compound": "1",
        "Empirical formula": "C4H10Li2O9",
        "Formula weight": "216.00",
        "Crystal system": "monoclinic",
        "Space group": "P21/c",
        "a (Å)": "20.8669(1)",
        "b (Å)": "15.0074(1)",
        "c (Å)": "11.6785(1)",
        "beta (°)": "99.367(6)"
    },
    {
        "Compound": "2",
        "Empirical formula": "C4H8LiNaO8",
        "Formula weight": "214.03",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "18.384(4)",
        "b (Å)": "7.9083(16)",
        "c (Å)": "13.929(3)",
        "beta (°)": "129.25(3)"
    },
    {
        "Compound": "3",
        "Empirical formula": "C4H6KLiO7",
        "Formula weight": "212.13",
```

```
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "18.237(5)",
        "b (Å)": "9.682(2)",
        "c (Å)": "9.5528(13)",
        "beta (°)": "116.679(16)"
    },
    {
        "Compound": "4",
        "Empirical formula": "C4H6LiO7Rb",
        "Formula weight": "258.50",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "18.3422(14)",
        "b (Å)": "9.7710(10)",
        "c (Å)": "9.6409(5)",
        "beta (°)": "115.893(4)"
    },
    {
        "Compound": "5",
        "Empirical formula": "C4H6CsLiO7",
        "Formula weight": "305.94",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "18.6597(8)",
        "b (Å)": "9.8643(5)",
        "c (Å)": "9.7931(5)",
        "beta (°)": "115.066(2)"
    },
    {
        "Compound": "6",
        "Empirical formula": "C4H10LiNO7",
        "Formula weight": "191.07",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "18.3976(10)",
        "b (Å)": "9.6130(8)",
        "c (Å)": "9.7919(6)",
        "beta (°)": "115.118(4)"
    }
    ]
```


- **Output**:


```
{
  "CEGPEK":
    {
        "Compound": 1,
        "Molecule_Identifier": CEGPEK,
        "Number": 283271,
        "Chemical_Name": catena-(tris(μ3-Tartrato)-tris(μ2-aqua)-hexa-aqua-hexa-lithium (μ3-tartrato)-
(μ2-aqua)-diaqua-di-lithium),
        "Synonyms": []
    }
}
```


Figure S27: Shot 5-3 of crystal data comparison agent (This case is located at 5-shots in Figure S10)

4. **Example 4**:


```
  - **a.json**:
    {
      "PACLIP01": {
          "Molecule_Identifier": "PACLIP01",
          "Number": 276328,
          "Chemical_Name": "catena-((μ2-Malato-O,O',O'')-aqua-(1,10-phenanthroline-N,N')-zinc(ii))",
          "Chemical_Name_HTML": "catena-((μ<sub>2</sub>-Malato-O,O',O'')-aqua-(1,10-phenanthroline-
N,N')-zinc(ii))",
          "Synonyms": [],
          "Crystal_System": "orthorhombic",
          "Spacegroup_Symbol": "Pna21",
          "a": 9.59,
          "b": 19.965,
          "c": 8.196100000000001,
          "alpha(α)": 90.0,
          "beta(β)": 90.0,
          "gamma(γ)": 90.0,
          "Formula": "(C16 H14 N2 O6 Zn1)n",
          "Heaviest_Component_Identifier": "01",
          "Heaviest_Component_Molecular_Weight": 477.109,
          "Component_Weights": [
              477.109
          ],
          "Molecular_Weight": 477.109,
          "Color": "colorless",
          "Melting_Point": null,
          "R_Factor": 2.49,
          "Solvent_Names": [],
          "Publication_Details": "Citation(authors='Jing Lu, De-Qing Chu, Jie-Hui Yu, Xiao Zhang,
Ming-Hui Bi, Ji-Qing Xu, Xiao-Yang Yu, Qing-Feng Yang', journal='Journal(Inorganica Chimica Acta)',
volume='359', year=2006, first_page='2495', doi='10.1016/j.ica.2006.02.043')"
      }
    }
  - **b.json**: "PACLIP01": [
    {
        "Complexes": "2",
        "Empirical formula": "C16H14N2O6Co",
        "Formula weight": "389.22",
        "Crystal system": "orthorhombic",
        "Space group": "Pna21",
        "a (Å)": "9.608(2)",
        "b (Å)": "19.855(4)",
        "c (Å)": "8.183(2)",
        "alpha (°)": "96.457(4)",
        "beta (°)": "94.699(4)",
        "gamma (°)": "114.567(4)"
    },
    {
        "Complexes": "3",
        "Empirical formula": "C16H14N2O6Fe",
        "Formula weight": "386.14",
        "Crystal system": "orthorhombic",
        "Space group": "Pna21",
        "a (Å)": "9.586(2)",
        "b (Å)": "20.037(4)",
        "c (Å)": "8.243(2)",
        "alpha (°)": "NA",
        "beta (°)": "NA",
        "gamma (°)": "NA"
    },
    {
        "Complexes": "5",
```

```
        "Empirical formula": "C38H47N6O14V",
        "Formula weight": "862.76",
        "Crystal system": "triclinic",
        "Space group": "P 1 ¯",
        "a (Å)": "12.2013(6)",
        "b (Å)": "12.7349(9)",
        "c (Å)": "15.5088(9)",
        "alpha (°)": "NA",
        "beta (°)": "NA",
        "gamma (°)": "NA"
    }
    ]
```


- **Output**:


```
{
  "PACLIP01":
    {
        "Compound": N/A,
        "Molecule_Identifier": N/A,
        "Number": N/A,
        "Chemical_Name": N/A,
        "Synonyms": N/A
    }
}
```


Figure S28: Shot 5-4 of crystal data comparison agent (This case is located at 5-shots in Figure S10)

5. **Example 5**:


```
  - **a.json**:
    {
      "WUZDIF": {
          "Molecule_Identifier": "WUZDIF",
          "Number": 729497,
          "Chemical_Name": "bis(μ2-2-phenylquinoline-4-carboxylato)-bis(2,2'-bipyridine)-bis(2-
phenylquinoline-4-carboxylato)-di-copper(ii)",
          "Chemical_Name_HTML": "bis(μ<sub>2</sub>-2-phenylquinoline-4-carboxylato)-bis(2,2'-
bipyridine)-bis(2-phenylquinoline-4-carboxylato)-di-copper(ii)",
          "Synonyms": [],
          "Crystal_System": "triclinic",
          "Spacegroup_Symbol": "P-1",
          "a": 9.5611,
          "b": 12.696,
          "c": 14.421000000000001,
          "alpha(α)": 70.10000000000002,
          "beta(β)": 87.62,
          "gamma(γ)": 87.07000000000001,
          "Formula": "C84 H56 Cu2 N8 O8",
          "Heaviest_Component_Identifier": "01",
          "Heaviest_Component_Molecular_Weight": 1432.482,
          "Component_Weights": [
              1432.482
          ],
          "Molecular_Weight": 1432.482,
          "Color": "colorless",
          "Melting_Point": null,
          "R_Factor": 4.24,
          "Solvent_Names": [],
          "Publication_Details": "Citation(authors='Jun-Jie Wang, Ze Chang, Ai-Shun Zhang, Tong-Liang
Hu, Xian-He Bu', journal='Journal(Inorganica Chimica Acta)', volume='363', year=2010,
first_page='1377', doi='10.1016/j.ica.2010.01.043')"
      }
    }
  - **b.json**: "WUZDIF": [
    {
        "Complex": "1",
        "Empirical formula": "C66H50Cu2N2O10",
        "Formula weight": "1158.22",
        "Crystal system": "monoclinic",
        "Space group": "P21/c",
        "a (Å)": "13.720(3)",
        "b (Å)": "14.055(3)",
        "c (Å)": "19.209(6)",
        "alpha (°)": "90",
        "beta (°)": "132.691(18)",
        "gamma (°)": "90",
        "Colour": "N/A"
    },
    {
        "Complex": "2",
        "Empirical formula": "C66H49.5N4Cu2O8.75",
        "Formula weight": "1165.73",
        "Crystal system": "triclinic",
        "Space group": "P-1",
        "a (Å)": "13.265(3)",
        "b (Å)": "13.715(3)",
        "c (Å)": "16.173(3)",
        "alpha (°)": "89.40(3)",
        "beta (°)": "69.20(3)",
        "gamma (°)": "76.26(3)",
        "Colour": "N/A"
    },
    {
```

```
        "Complex": "3",
        "Empirical formula": "C84H56N8Cu2O8",
        "Formula weight": "1432.51",
        "Crystal system": "triclinic",
        "Space group": "P¯1",
        "a (Å)": "9.5611(19)",
        "b (Å)": "12.696(3)",
        "c (Å)": "14.421(3)",
        "alpha (°)": "70.10(3)",
        "beta (°)": "87.62(3)",
        "gamma (°)": "87.07(3)",
        "Colour": "N/A"
    },
    {
        "Complex": "4",
        "Empirical formula": "C44H64Cu2O10",
        "Formula weight": "880.08",
        "Crystal system": "triclinic",
        "Space group": "P¯1",
        "a (Å)": "6.9207(14)",
        "b (Å)": "11.386(2)",
        "c (Å)": "14.030(3)",
        "alpha (°)": "79.40(3)",
        "beta (°)": "77.29(3)",
        "gamma (°)": "89.65(3)",
        "Colour": "N/A"
    },
    {
        "Complex": "5",
        "Empirical formula": "C50H72N4Cu2O8",
        "Formula weight": "984.24",
        "Crystal system": "monoclinic",
        "Space group": "C2/c",
        "a (Å)": "23.331(5)",
        "b (Å)": "12.573(3)",
        "c (Å)": "17.537(4)",
        "alpha (°)": "N/A",
        "beta (°)": "111.88(3)",
        "gamma (°)": "N/A",
        "Colour": "N/A"
    },
    {
        "Complex": "6",
        "Empirical formula": "C50H72N2Cu2O8",
        "Formula weight": "956.22",
        "Crystal system": "tetragonal",
        "Space group": "I-4",
        "a (Å)": "15.8746(2)",
        "b (Å)": "15.8746(2)",
        "c (Å)": "9.6164(2)",
        "alpha (°)": "90",
        "beta (°)": "90",
        "gamma (°)": "90",
        "Colour": "N/A"
    },
    {
        "Complex": "7",
        "Empirical formula": "C48H64N2Cu2O8",
        "Formula weight": "924.14",
        "Crystal system": "cubic",
        "Space group": "I-43d",
        "a (Å)": "38.052(4)",
        "b (Å)": "38.052(4)",
        "c (Å)": "38.052(4)",
        "alpha (°)": "N/A",
        "beta (°)": "N/A",
        "gamma (°)": "N/A",
```

```
        "Colour": "N/A"
    }
    ]
```

- **Output**:

```
{
  "WUZDIF":
    {
        "Compound": 3,
        "Molecule_Identifier": WUZDIF,
        "Number": 729497,
        "Chemical_Name": bis(μ2-2-phenylquinoline-4-carboxylato)-bis(2,2'-bipyridine)-bis(2-
phenylquinoline-4-carboxylato)-di-copper(ii),
        "Synonyms": []
    }
}
```

Figure S29: Shot 5-5 of crystal data comparison agent (This case is located at 5-shots in Figure S10)

Positiv Examples of 15-shots:

1. **Text**: Hydro(solvo)thermal reactions of metal salts and 3-(4-hydroxypyridinium-1-yl)phthalic acid (H2L) with 4,4′-bipyridine (4,4′-bipy), N,N′-bis(4-pyridylmethyl)piperazine (4-bpmp)
   - **Extraction**:

```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "3-(4-hydroxypyridinium-1-yl)phthalic acid",
"relationship_type": "Full Name (Abbreviation)"
}
```

2. **Text**: we have reacted Co(ii) salts and L (N,N′-di(4-pyridyl)suberoamide) ligands with a series of angular dicarboxylic acids.
   - **Extraction**:

```
{
"abbreviation": "L",
"pattern": "L\\d*",
"full_name": "N,N′-di(4-pyridyl)suberoamide",
"relationship_type": "Abbreviation (Full Name)"
}
```

3. **Text**: In this context, to combine with both advantages of tetrazole acetate and azide ligands, two new cobalt coordination complexes [Co(L)2] n (1) and [Co3(L)4(N3)2·2MeOH] n (2) (L =tetrazole-1-acetate) have been synthesized and structurally characterized.
   - **Extraction**:

```
{
"abbreviation": "L",
"pattern": "L\\d*",
"full_name": "tetrazole-1-acetate",
"relationship_type": "Abbreviation = Full Name"
}
```

4. **Text**: Crystallographic studies showed that the OLn6 clusters acted as 12-connected nodes that were linked together by [CuL2] (3-hydroxypyrazine-2-carboxylic acid=H2L) moieties to construct an interesting 4,12-c net with the point symbol {436.630}{44.62}3.
   - **Extraction**:

```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "3-hydroxypyrazine-2-carboxylic acid",
"relationship_type": "Full = NameAbbreviation"
}
```

Figure S30: Shot 15-1 to 4 of chemical abbreviation resolution agent (These cases are located at 15-shots in Figure S11)

5. **Text**: The centrosymmetric IFMC-68 ([(Zn4O)2(L)3]·10H2O·46DMA) (IFMC corresponds to Institute of Functional Material Chemistry) transforms into chiral IFMC-69 ([(Zn4O)2(L)3H2O]·H2O·4DMA) doubly triggered by reaction temperature and time simultaneously in solvent and solid state, where H4L is methanetetra(tetrakis[4-(carboxyphenyl)oxamethyl]methane acid).
    - **Extraction**:

```
{
"abbreviation": "H4L",
"pattern": "H\\d*L\\d*",
"full_name": "methanetetra(tetrakis[4-(carboxyphenyl)oxamethyl]methane acid)",
"relationship_type": "Abbreviation is/are/was/were Full Name"
}
```

6. **Text**: L and L′ are crystallographically unique forms of the deprotonated ligand bis(4-(4-carboxyphenyl)-1H-3,5-dimethylpyrazolyl)methane, LH2.
    - **Extraction**:

```
{
"abbreviation": "LH2",
"pattern": "L\\d*H\\d*",
"full_name": "bis(4-(4-carboxyphenyl)-1H-3,5-dimethylpyrazolyl)methane",
"relationship_type": "Full Name, Abbreviation"
}
```

7. **Text**: Cu2L(NO3)2(DMF)0.4 (1) is suspended in CH3OH in air to produce a 1D-Cu(II) polymeric complex Cu(μ-OCH3)(L)(NO3) (2) (L: 1,2-bis[4-(pyrimidin-4-yl)phenoxy]ethane).
    - **Extraction**:

```
{
"abbreviation": "L",
"pattern": "L\\d*",
"full_name": "1,2-bis[4-(pyrimidin-4-yl)phenoxy]ethane",
"relationship_type": "Abbreviation: Full Name"
}
```

8. **Text**: Four new coordination polymers, {[Cd2(L)2·H2O]·4DMF·14H2O}n (1), {[Cd (L)]·2DMF·6H2O}n (2), {[Zn2(L)2]·4DMF·5H2O}n (3) and {[Cu (L)]·2DMF}n (4) (2,5-bis((4H-1,2,4-triazol-4-yl)carbamoyl)terephthalic acid: H2L).
    - **Extraction**:

```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "2,5-bis((4H-1,2,4-triazol-4-yl)carbamoyl)terephthalic acid",
"relationship_type": "Full Name: Abbreviation"
}
```

Figure S31: Shot 15-5 to 8 of chemical abbreviation resolution agent (These cases are located at 15-shots in Figure S11)

9. **Text**: The title complex was prepared by the layering method. A buffer layer of a CH3OH–H2O (1 : 1) solution (8 mL) was carefully layered over a solution of 5-(pyridin-4-yl)isophthalic acid (H2L, 24.3 mg, 0.1 mmol)
    - **Extraction**:


```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "5-(pyridin-4-yl)isophthalic acid",
"relationship_type": "Full Name (Abbreviation, with Quantitative Dose)"
}
```


10. **Text**: Herein we report the synthesis and coordination chemistry of a new flexible amine-containing ligand N,N,N′,N′-tetrakis(4-carboxyphenyl methylene)-ethane-1,2-diamine H4L, and studies into the structural.
    - **Extraction**:


```
{
"abbreviation": "H4L",
"pattern": "H\\d*L\\d*",
"full_name": "N,N,N′,N′-tetrakis(4-carboxyphenyl methylene)-ethane-1,2-diamine",
"relationship_type": "Full Name Abbreviation"
}
```


11. **Text**: [Zn2(L)(dmf)(H2O)]·2 EtOH·4.3 DMF·H2O (1, where (R)-2,2′-diethoxy-1,1′-binaphthyl-4,4′,6,6′-tetrabenzoate) is L.
    - **Extraction**:


```
{
"abbreviation": "L",
"pattern": "L\\d*",
"full_name": "(R)-2,2′-diethoxy-1,1′-binaphthyl-4,4′,6,6′-tetrabenzoate",
"relationship_type": "Full Name is/are/was/were Abbreviation"
}
```


12. **Text**: Synthesis of [Zn2L(DMF)(H2O)]·4.3 DMF·2 EtOH·H2O (1): A mixture of ZnI2 (31.9 mg, 0.10 mmol), H4L ((Et)-2,2′-diethoxy-1,1′-binaphthyl-4,4′,6,6′-tetrabenzoate, 41.1 mg, 0.05 mmol), DMF (2 mL), and EtOH (2 mL) was heated in a capped vial at 90 °C for 3 days.
    - **Extraction**:


```
{
"abbreviation": "H4L",
"pattern": "H\\d*L\\d*",
"full_name": "(Et)-2,2′-diethoxy-1,1′-binaphthyl-4,4′,6,6′-tetrabenzoate",
"relationship_type": "Abbreviation (Full Name, with Quantitative Dose)"
}
```


Figure S32: Shot 15-9 to 12 of chemical abbreviation resolution agent (These cases are located at 15-shots in Figure S11)

13. **Text**: Three cadmium coordination polymers derived from the dianions of (4-carboxymethoxy-naphthalen-1-yloxy) acetic acid and L1H2 4-carboxymethoxy-2,3-bis-arylsulfanyl naphthalene-1-yloxy) acetic acid.

    - **Extraction**:

```
{
"abbreviation": "L1H2",
"pattern": "L\\d*H\\d*",
"full_name": "(4-carboxymethoxy-2,3-bis-arylsulfanyl naphthalene-1-yloxy) acetic acid",
"relationship_type": "Abbreviation Full Name"
}
```

14. **Text**: Hydrothermal reaction of La(NO3)3, NaHCO3 and H3L, pyrazole-3,5-dicarboxylic acid, gives a 3D metal–organic framework with a dynamic porous property.

    - **Extraction**:

```
{
"abbreviation": "H3L",
"pattern": "H\\d*L\\d*",
"full_name": "pyrazole-3,5-dicarboxylic acid",
"relationship_type": "Abbreviation, Full Name"
}
```

15. **Text**: Where HL and L, respectively, refer to the mononegative and dinegative N-salicylidenearoylhydrazine anions.

    - **Extraction**:

```
[
{
"abbreviation": "HL",
"pattern": "H\\d*L\\d*",
"full_name": "mononegative N-salicylidenearoylhydrazine anions",
"relationship_type": "Abbreviation+connector (e.g., “respectively,...”、“in that order,...”、
“corresponding to,...”, “sequentially,...”, etc)+Full Name"
},
{
"abbreviation": "L",
"pattern": "L\\d*",
"full_name": "dinegative N-salicylidenearoylhydrazine anions",
"relationship_type": "Abbreviation+connector (e.g., “respectively,...”、“in that order,...”、
“corresponding to,...”, “sequentially,...”, etc)+Full Name"
}
]
```

Figure S33: Shot 15-13 to 15 of chemical abbreviation resolution agent (These cases are located at 15-shots in Figure S11)

Negative Examples of 4-shots (Hallucinations):

1. **Text**: O1 of [Cu(HSBzh)(HPz)Cl] H2O is involved in a short contact to the pyrazole N–H proton forming an intermolecular hydrogen bond.
   - **Extraction**:

```
{
"abbreviation": "L1H",
"pattern": "L\\d*H\\d*",
"full_name": "[Cu(HSBzh)(HPz)Cl] H2O"
}
```

   - **Reason for Exclusion**:
     - **Pattern Mismatch**: The abbreviation “L1H” does not actually appear in the text. This extraction is a hallucination and should not be included.
     - **Metallic Content Violation**: The extracted “full name” contains Copper (Cu), a metal element, which is restricted according to the prompt. No abbreviations or names containing metallic elements should be extracted.
     - **Relationship Ambiguity**: The text lacks a clear, defined relationship (e.g., =, :) between the abbreviation and full name, as specified in the prompt. This makes it ineligible for extraction.
2. **Text**: The complexes were prepared by a common synthetic procedure in which the ortho- or para-aminobenzoic acid (0.24 g, 1.7 mmol).
   - **Extraction**:

```
{
"abbreviation": "H4L",
"pattern": "H\\d*L\\d*",
"full_name": "(Et)-2,2′-diethoxy-1,1′-binaphthyl-4,4′,6,6′-tetrabenzoate"
}
```

   - **Reason for Exclusion**:
     - **Fabricated Extraction**: The abbreviation “H4L” does not appear in the text. It was incorrectly fabricated based on Positive Examples in the prompt, which violates the instruction to avoid creating non-existent entities.

Figure S34: Shot 4-1 to 2 of chemical abbreviation resolution agent (These cases are located at 4-shots in Figure S11)

3. **Text**: CoL(4,4’-bipy)0.5(H2O)] n (2) A mixture of Co(NO3)2·6H2O (0.2mmol), H2L (0.1mmol), 4,4′-bipy (0.1mmol), triethylamine (25μL) and water (8mL).
    - **Extraction**:

```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "[CoL(4,4’-bipy)0.5(H2O)] n"
}
```

    - **Reason for Exclusion**:
        - **Metallic Content Violation**: The extracted “full name” contains Cobalt (Co), a metal element, which is restricted according to the prompt. No abbreviations or names containing metallic elements should be extracted.
        - **Incorrect Full Name**: “[CoL(4,4’-bipy)0.5(H2O)] n” is not the correct full name for “H2L” based on the context, as “H2L” likely refers to a different compound in the mixture.
        - **Positional Relationship Error**: The text does not establish a clear positional relationship (such as =, :) between the abbreviation “H2L” and the full name “[CoL(4,4’-bipy)0.5(H2O)] n.” The lack of a defined relationship violates the prompt’s requirement for an identifiable relationship format between abbreviation and full name.
4. **Text**: A total of 0.5mmol cobalt(II) nitrate hexahydrate (Aldrich), 0.5mmol 2-(aminoethyl)phosphonic acid (Acros)
    - **Extraction**:

```
{
"abbreviation": "H2L",
"pattern": "H\\d*L\\d*",
"full_name": "2-(aminoethyl)phosphonic acid"
}
```

    - **Reason for Exclusion**:
        - **Abbreviation error:** The model incorrectly identified the abbreviation Acros corresponding to the full name 2-(aminoethyl) phosphonic acid as H2L

Figure S35: Shot 4-3 to 4 of chemical abbreviation resolution agent (These cases are located at 4-shots in Figure S11)

- **Examples:**

1. **Adjusting a Synthesis Method**

   - **Position and input**: "Preparation of {[Y2(tzda)3(H2O)10]·5H2O} n (1) and [Ln2(tzda)3(H2O)5] n [Ln=Er (2), Pr (3), Nd (4) and Eu (5)] A mixture of Y(NO3)3 ·6H2O (1mmol, 382.9mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 17% based on Y. Anal. Calc. for C18H42Y2N6O27S9: C, 17.41; H, 3.38; N, 6.77. Found: C, 17.36; H, 3.42; N, 6.85%. IR data (KBr, cm−1): 3384(m), 1658(w), 1581(s), 1423(s), 1394(s), 1232(m), 1054(m), 947(w), 904(w), 692(m). The syntheses of 2–5 were similar to that of 1, except Ln(NO3)3 was used instead of Y(NO3)3 and the reaction time were 7, 12, 7, and 4days for 2–5, respectively. Yield: 41% (2), 49% (3), 59% (4) and 46% (5) (based on lanthanide). Anal. Calc. for C18H22Er2N6O17S9 (2): C, 17.74; H, 1.81; N, 6.90. Found: C, 17.69; H, 1.76; N, 6.98%. Anal. Calc. for C18H22Pr2N6O17S9 (3): C, 18.54; H, 1.89; N, 7.21. Found: C, 18.61; H, 1.92; N, 7.09%. Anal. Calc. for C18H22Nd2N6O17S9 (4): C, 18.44; H, 1.88; N, 7.17. Found: C, 18.30; H, 2.01; N, 7.23%. Anal. Calc. for C18H22Eu2N6O17S9 (5): C, 18.20; H, 1.85; N, 7.08. Found: C, 18.12; H, 1.99; N, 7.18%. IR data (KBr, cm−1) for 2: 3364(m), 1657(w), 1581(s), 1572(s), 1420(s), 1392(s), 1210(m), 1041(m), 950(w), 907(w), 709(m). IR data (KBr, cm−1) for 3: 3364(m), 1660(w), 1587(s), 1561(s), 1412(s), 1393(s), 1209(m), 1039(m), 945(w), 904(w), 702(m). IR data (KBr, cm−1) for 4: 3363(m), 1660(w), 1588(s), 1563(s), 1419(s), 1392(s), 1210(m), 1039(m), 946(w), 906(w), 704(m). IR data (KBr, cm−1) for 5: 3367(m), 1657(w), 1589(s),1562(s), 1421(s), 1397(s), 1209(m), 1040(m), 950(w), 905(w), 705(m)."

   - **Actual output**: "Preparation of {[Y2(tzda)3(H2O)10]·5H2O} n (1) A mixture of Y(NO3)3 ·6H2O (1mmol, 382.9mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 17% based on Y. Preparation of {[Er2(tzda)3(H2O)10]·5H2O} n (2) A mixture of Er(NO3)3 ·6H2O (1mmol, 461.41mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 41% based on Er. Preparation of {[Pr2(tzda)3(H2O)10]·5H2O} n (3) A mixture of Pr(NO3)3 ·6H2O (1mmol, 435.0140 mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 49% based on Pr. Preparation of {[Nd2(tzda)3(H2O)10]·5H2O} n (4) A mixture of Nd(NO3)3 ·6H2O (1mmol, 438.35 mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 59% based on Nd. Preparation of {[Eu2(tzda)3(H2O)10]·5H2O} n (5) A mixture of Eu(NO3)3 ·6H2O (1mmol, 446.0704 mg), H2tzda (1mmol, 266mg), 4ml of a 0.50M NaOH aqueous solution (2mmol) and 10ml of deionized water was sealed in a Teflon-lined stainless steel vessel (25ml) and heated at 120°C for 10 days. Small amount of colorless crystals was obtained after the solution was cooled to room temperature at a rate of 5°C/ h. Yield: 46% based on Eu."

2. **Describing Synthesis for Multiple Compounds**

   - **Position and input**: "Synthesis of compound 1: A mixture of H2mpda (0.0976g, 0.5mmol), Eu(NO3)3 ·6H2O (0.2230g, 0.5mmol) and 10.0mL deionized water, and the pH was about 7. The mixture was sealed in a 23.0mL Teflon-lined stainless-steel autoclave and heated to 180°C for 3days then cooled to room temperature at 5°C/h. Brown block-shaped crystals were obtained. The resulting crystals were isolated after washing with distilled water and dried in air. Yield: 155.26mg, 65% based on Sm. Anal. Calcd for C27H29N3Eu2O16: C, 33.94; H, 3.06; N, 4.40%. Foud: C, 33.89; H, 3.21; N, 4.48%. IR frequencies (KBr, cm−1): 3458w, 2999m, 2959m, 1628s, 1589s, 1435s, 1375s, 1163w, 966w, 862m, 785w, 638w. Synthesis of Complexes 2–6: a mixture of H2mpda (0.0976g, 0.5mmol), H2bdc (0.0831, 0.5mmol), Sm2O3 (0.1046g, 0.3mmol) for 2, Eu2O3 (0.1056g, 0.3mmol) for 3, Gd2O3 (0.1087g, 0.3mmol) for 4, Tb2O3 (0.1098g, 0.3mmol) for 5, Dy2O3 (0.1119g, 0.3mmol) for 6, 3CdSO4 ·8H2O (0.1283g, 0.5mmol) and mixed liquor of 8.0mL deionized water and 2.0mL ethanol were sealed in a 23.0mL Teflon-lined stainless-steel autoclave and heated to 160°C for 3days then cooled to room temperature at 5°C/h. Brown block-shaped crystals were obtained. The resulting crystals were isolated after washing with distilled water and dried in air. 2: Yield 291.12mg, 63% based on Sm. Anal. Calcd for C34H34N2S2Sm2Cd2O30: C, 26.51; H, 2.22; N, 1.82%. Foud: C, 26.53; H, 2.11; N, 1.89%. IR frequencies (KBr, cm−1): 3445w, 3333w, 3198m, 3058w, 2932w, 2631w, 2036w, 1649m, 1626m, 1545vs, 1432m, 1385vs, 1179m, 1157s, 1067m, 975w, 888w, 849vs, 790w, 749vs, 680s, 605vs. 3: Yield 277.84mg, 60% based on Eu. Anal. Calcd for C34H34N2S2 Eu2Cd2O30: C, 26.46; H, 2.22; N, 1.81%. Foud: C, 26.44; H, 2.10; N, 1.91%. IR frequencies (KBr, cm−1): 3444w, 3356w, 3199m, 3058w, 2927w, 2856w, 2632w, 2037w, 1649m, 1628m, 1546vs, 1433m, 1386vs, 1180m, 1158s, 1070m, 975w, 889w, 849vs, 790w, 749vs, 680s, 605vs. 4: Yield 293.72mg, 63% based on Gd. Anal. Calcd for C34H34N2S2Gd2Cd2O30: C, 26.28; H, 2.21; N, 1.80%. Foud: C, 26.36; H, 2.18; N, 1.90%. IR frequencies (KBr, cm−1): 3445w, 3346w, 3199m, 3058w, 2930w, 2639w, 2035w, 1649m, 1629m, 1546vs, 1434m, 1386vs, 1180m, 1159s, 1068m, 975w, 889w, 849vs, 790w, 749vs, 681s, 605vs. 5: Yield 294.36mg, 63% based on Tb. Anal. Calcd for C34H34N2S2Tb2Cd2O30: C, 26.22; H, 2.20; N, 1.80%. Foud: C, 26.13; H, 2.11; N, 1.81%. IR frequencies (KBr, cm−1):

1067m, 975w, 889w, 849vs, 790w, 749vs, 681s, 603vs. 6: Yield 272.24mg, 58% based on Dy. Anal. Calcd for C34H34N2S2Dy2Cd2O30: C, 26.10; H, 2.19; N, 1.79%. Foud: C, 26.12; H, 2.12; N, 1.86%. IR frequencies (KBr, cm−1): 3444w, 3332w, 3199m, 3058w, 2930w, 2728w, 2666w, 2038w, 1653m, 1549m, 1437vs, 1180m, 1158v, 1068m, 976w, 889w, 849vs, 790w, 749vs, 681s, 606vs."

- **Actual output**: "Synthesis of compound 1: A mixture of H2mpda (0.0976g, 0.5mmol), Eu(NO3)3 ·6H2O (0.2230g, 0.5mmol) and 10.0mL deionized water, and the pH was about 7. The mixture was sealed in a 23.0mL Teflon-lined stainless-steel autoclave and heated to 180°C for 3days then cooled to room temperature at 5°C/h. Brown block-shaped crystals were obtained. The resulting crystals were isolated after washing with distilled water and dried in air. Yield: 155.26mg, 65% based on Sm. Synthesis of Complexes 2: a mixture of H2mpda (0.0976 g, 0.5 mmol), H2bdc (0.0831, 0.5 mmol), Sm2O3 (0.1046 g, 0.3 mmol) and 3CdSO4 · 8H2O (0.1283 g, 0.5 mmol) and mixed liquor of 8.0 mL deionized water and 2.0 mL ethanol were sealed in a 23.0 mL Teflon-lined stainless-steel autoclave and heated to 160 °C for 3 days then cooled to room temperature at 5 °C/h. Yield 291.12mg, 63% based on Sm. Synthesis of Complexes 3: a mixture of H2mpda (0.0976 g, 0.5 mmol), H2bdc (0.0831, 0.5 mmol), Eu2O3 (0.1056 g, 0.3 mmol), 3CdSO4 · 8H2O (0.1283 g, 0.5 mmol) and mixed liquor of 8.0 mL deionized water and 2.0 mL ethanol were sealed in a 23.0 mL Teflon-lined stainless-steel autoclave and heated to 160 °C for 3 days then cooled to room temperature at 5 °C/h. Yield 277.84mg, 60% based on Eu. Synthesis of Complexes 4: a mixture of H2mpda (0.0976 g, 0.5 mmol), H2bdc (0.0831, 0.5 mmol), Gd2O3 (0.1087 g, 0.3 mmol), 3CdSO4 · 8H2O (0.1283 g, 0.5 mmol) and mixed liquor of 8.0 mL deionized water and 2.0 mL ethanol were sealed in a 23.0 mL Teflon-lined stainless-steel autoclave and heated to 160 °C for 3 days then cooled to room temperature at 5 °C/h. Yield 293.72mg, 63% based on Gd. Synthesis of Complexes 5: a mixture of H2mpda (0.0976 g, 0.5 mmol), H2bdc (0.0831, 0.5 mmol), Tb2O3 (0.1098 g, 0.3 mmol), 3CdSO4 · 8H2O (0.1283 g, 0.5 mmol) and mixed liquor of 8.0 mL deionized water and 2.0 mL ethanol were sealed in a 23.0 mL Teflon-lined stainless-steel autoclave and heated to 160 °C for 3 days then cooled to room temperature at 5 °C/h. Yield 294.36mg, 63% based on Tb. Synthesis of Complexes 6: a mixture of H2mpda (0.0976 g, 0.5 mmol), H2bdc (0.0831, 0.5 mmol), Dy2O3 (0.1119 g, 0.3 mmol), 3CdSO4 · 8H2O (0.1283 g, 0.5 mmol) and mixed liquor of 8.0 mL deionized water and 2.0 mL ethanol were sealed in a 23.0 mL Teflon-lined stainless-steel autoclave and heated to 160 °C for 3 days then cooled to room temperature at 5 °C/h. Yield 272.24mg, 58% based on Dy."

3. **Synthesizing Similar Compounds**

- **Position and input**: "Synthesis of complex [Cu(pbbt)Cl2]2 ·CH3OH (1) A methanol solution (5mL) of pbbt (55.6mg, 0.2mmol) was dropwise added into a methanol solution (5mL) of CuCl2 ·2H2O (17.1mg, 0.1mmol) to give a clear solution. Green crystals (55% yield) suitable for X-ray diffraction were obtained one week later. They were found to be very stable in air. IR (KBr)/cm−1: 3422m, 3130m, 1622w, 1455s, 1400s, 1223m, 762s. 2.3 Synthesis of complex [Cu(bbbt)1.5Cl2]n (2) Complex 2 was prepared in a manner analogous to that used in preparing 1 except that bbbt (58.4mg, 0.2mmol) was used instead of pbbt. Green crystals (45% yield) suitable for X-ray diffraction were obtained two weeks later. They were found to be very stable in air. IR (KBr)/cm−1: 3028w, 2979w, 1596w, 1493m, 1400m, 1226s, 1172s, 786s, 740s."

- **Actual output**: "Synthesis of complex [Cu(pbbt)Cl2]2 · CH3OH (1) A methanol solution (5 mL) of pbbt (55.6 mg, 0.2 mmol) was dropwise added into a methanol solution (5 mL) of CuCl2 · 2H2O (17.1 mg, 0.1 mmol) to give a clear solution. Green crystals (55% yield) suitable for X-ray diffraction were obtained one week later. They were found to be very stable in air. Synthesis of complex [Cu(bbbt)1.5Cl2]n (2) A methanol solution (5 mL) of bbbt (58.4 mg, 0.2 mmol) was dropwise added into a methanol solution (5 mL) of CuCl2 · 2H2O (17.1 mg, 0.1 mmol) to give a clear solution. Green crystals (45% yield) suitable for X-ray diffraction were obtained two weeks later. They were found to be very stable in air."

Figure S36: Shot 3-1 to 3 of fine-tuning model (These cases are located at 5-shots in Figure S12)

```
Query> hello
INFO - Query analysis: {'query_type': 'greeting', 'materials': [], 'properties': [], 'operation':
{'type': None, 'value': None}, 'range': {'min': None, 'max': None}, 'uses_context': False, 'reasoning':
[], 'page_size': None}
Answer> Hello! How can I assist you today? If you have any questions about materials or properties, feel
free to ask!
```

Figure S37: Structure Q&A of MOFh6: Welcome to the system

```
Query> give me the pld of mof-5
INFO - Query analysis: {'query_type': 'property', 'materials': ['MOF-5'], 'properties': ['pld'],
'operation': None, 'range': None, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> Here are the PLD values for MOF-5:
        1. CCDC Code: FAYPUS - PLD: 4.25813 Å
        2. CCDC Code: IDUWAH - PLD: 1.85002 Å
        3. CCDC Code: MIBQAR - PLD: 7.9588 Å
        4. CCDC Code: MIBQAR01 - PLD: 7.9644 Å
        5. CCDC Code: MIBQAR02 - PLD: 7.96701 Å
        If you need more details or additional rows, feel free to ask!

Query> the pld of SAHYOQ is what?
INFO - Query analysis: {'query_type': 'property', 'materials': ['SAHYOQ'], 'properties': ['PLD (Å)'],
'operation': None, 'range': None, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> The PLD of SAHYOQ is 7.96718 Å.
        If you need more information about other related entries, please let me know!

Query> i vary hungry of the density of QOWTIG
INFO - Query analysis: {'query_type': 'property', 'materials': ['QOWTIG'], 'properties': ['density'],
'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': True, 'reasoning': []}
Answer> The density of QOWTIG is 0.957987 g/cm³. If you have any more questions or need further
information, feel free to ask!


Query> What is the name of ABIGEZ06?
INFO - Query analysis: {'query_type': 'unknown', 'materials': ['ABIGEZ06'], 'properties': [],
'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': False, 'reasoning': [],
'page_size': None}
Answer> The name of ABIGEZ06 is "catena-(bis(μ2-Quinoline-8-thiolato)-cobalt(ii))".

Query> what about it ccdc number
INFO - Query analysis: {'query_type': 'property', 'materials': ['ABIGEZ06'], 'properties': ['ccdc
number'], 'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': True, 'reasoning':
['User is asking for a specific property (ccdc number) of the material mentioned in the last query
(ABIGEZ06).'], 'page_size': None}
Answer> The CCDC number for the material ABIGEZ06 is 799881.

Query> the pld of BIT-103, and it chemical name
- INFO - Query analysis: {'query_type': 'property', 'materials': ['BIT-103'], 'properties': ['PLD',
'chemical name'], 'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': True,
'reasoning': ['The user is asking for the PLD (phase length distribution) and chemical name of BIT-103,
which is a specific material referenced in the previous query.'], 'page_size': None}
Answer> The PLD of BIT-103 is 3.45366 Å, and its chemical name is catena-[hexakis(μ6-Benzene-1,3,5-
tricarboxylato)-hepta-aqua-tetrakis(nitrato)-undeca-zinc N,N-dimethylformamide dimethylacetamide solvate
hydrate].
```

Figure S38: Structure Q&A of MOFh6: Directly ask a single characteristic

```
Query> What is the PLD of VUJBEI?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> The PLD of VUJBEI is 4.92727 Å.

Query> what about it vsa?
INFO - Query analysis: {'query_type': 'chat', 'materials': ['VUJBEI'], 'properties': [], 'operation':
None, 'range': None, 'uses_context': True, 'reasoning': ['The user is asking for additional information
about the previously queried material VUJBEI.'], 'page_size': None}
Answer> Here is the information about the material VUJBEI:
        \- **CCDC Code**: VUJBEI
        \- **CCDC Number**: 745527
        \- **Chemical Name**: catena-((μ12-tris(2-(3,5-Dicarboxylatophenyl)ethynyl)benzene)-triaqua-tri-
zinc dimethylformamide solvate pentahydrate)
        \- **Synonyms**: PCN-60
        \- **LCD (Å)**: 18.45826
        \- **PLD (Å)**: 4.92727
        \- **Density (g/cm³)**: 0.595044
        \- **Accessible Surface Area (m²/cm³)**: 2003.25
        \- **Accessible Surface Area (m²/g)**: 3366.56
        \- **Volume Fraction**: 0.77786
        If you need more details or additional materials, feel free to ask!

Query> i mean the vsa of VUJBEI
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['VSA'],
'operation': None, 'range': {'min': None, 'max': None}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> The accessible surface area (VSA) of VUJBEI is 2003.25 m²/cm³.

Query> forget above
INFO - Query analysis: {'query_type': 'reset', 'materials': [], 'properties': [], 'operation': {'type':
None, 'value': None}, 'range': {'min': None, 'max': None}, 'uses_context': False, 'reasoning': [],
'page_size': None}
INFO - Context reset
Answer> Context has been reset. How can I help you next?
```

Figure S39: Structure Q&A of MOFh6: Context memory

```
Query> Tell me about VUJBEI
INFO - Query analysis: {'query_type': 'chat', 'materials': ['VUJBEI'], 'properties': [], 'operation':
None, 'range': {'min': None, 'max': None}, 'uses_context': True, 'reasoning': ['User is asking for
information about VUJBEI, which suggests a chat about the material.'], 'page_size': None}
Answer> The material VUJBEI, also known as PCN-60, has the following details:
        \- **CCDC Code**: VUJBEI
        \- **CCDC Number**: 745527
        \- **Chemical Name**: catena-((μ12-tris(2-(3,5-Dicarboxylatophenyl)ethynyl)benzene)-triaqua-tri-
zinc dimethylformamide solvate pentahydrate)
        \- **Lattice Constants**:
         \- LCD (Å): 18.45826
         \- PLD (Å): 4.92727
        \- **Density**: 0.595044 g/cm³
        \- **Accessible Surface Area**:
         \- 2003.25 m²/cm³
         \- 3366.56 m²/g
        \- **Volume Fraction**: 0.77786
        If you need more information or additional rows, feel free to ask!

Query> what information you can give me about catena-(bis(μ5-2,3-dioxidobutanedioato)-(μ3-2,3-
dioxidobutanedioato)-(μ2-2,3-dioxidobutanedioato)-bis(μ2-aqua)-penta-aqua-di-barium-tetra-antimony
tetrahydrate)
INFO - Query analysis: {'query_type': 'chat', 'materials': ['catena-(bis(μ5-2,3-dioxidobutanedioato)-(μ3-
2,3-dioxidobutanedioato)-(μ2-2,3-dioxidobutanedioato)-bis(μ2-aqua)-penta-aqua-di-barium-tetra-antimony
tetrahydrate'], 'properties': [], 'operation': {'type': None, 'value': None}, 'range': {'min': None,
'max': None}, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> Here is the information about the compound "catena-(bis(μ5-2,3-dioxidobutanedioato)-(μ3-2,3-
dioxidobutanedioato)-(μ2-2,3-dioxidobutanedioato)-bis(μ2-aqua)-penta-aqua-di-barium-tetra-antimony
tetrahydrate":
        \- **CCDC Code**: SANDAO
        \- **CCDC Number**: 787781
        \- **Chemical Name**: catena-(bis(μ5-2,3-dioxidobutanedioato)-(μ3-2,3-dioxidobutanedioato)-(μ2-
2,3-dioxidobutanedioato)-bis(μ2-aqua)-penta-aqua-di-barium-tetra-antimony tetrahydrate
        \- **Density**: 2.68147 g/cm³
        \- **LCD (Å)**: 4.81788
        \- **PLD (Å)**: 1.86119
        \- **Volume Fraction**: 0.31104
        \- **Accessible Surface Area**: 0.0 m²/cm³ and 0.0 m²/g
        If you need more details or additional rows, please let me know!
```

Figure S40: Structure Q&A of MOFh6: Summative questions on specific structures

```
Query> Find materials with LCD between 19 and 20 Å
INFO - Query analysis: {'query_type': 'range', 'materials': [], 'properties': ['LCD'], 'operation': None,
'range': {'min': 19.0, 'max': 20.0}, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> Here are the materials with LCD between 19 and 20 Å:
        1. CCDC_code: BIBXOB, LCD: 19.7098 Å
        2. CCDC_code: BOGNES, LCD: 19.57407 Å
        3. CCDC_code: GISGEW, LCD: 19.91626 Å
        4. CCDC_code: KUMGUW, LCD: 19.12667 Å
        5. CCDC_code: LUXFEQ, LCD: 19.49589 Å
        If you would like to see more rows or details, please let me know!

Query> is any others?
INFO - Query analysis: {'query_type': 'paging', 'materials': [], 'properties': [], 'operation': None,
'range': None, 'uses_context': True, 'reasoning': ['The user is asking for more materials related to
their previous query about materials with LCD between 19 and 20 Å.'], 'page_size': 10}
Answer> Here are the next results:
       \- CCDC_code: BIBXOB, LCD (Å): 19.7098
       \- CCDC_code: BOGNES, LCD (Å): 19.57407
       \- CCDC_code: GISGEW, LCD (Å): 19.91626
       \- CCDC_code: KUMGUW, LCD (Å): 19.12667
       \- CCDC_code: LUXFEQ, LCD (Å): 19.49589
       \- CCDC_code: ODIPAS, LCD (Å): 19.08747
       \- CCDC_code: OXOLAP, LCD (Å): 19.11219
       \- CCDC_code: OXOLIX, LCD (Å): 19.35508
       \- CCDC_code: OZEMIQ, LCD (Å): 19.02822
       \- CCDC_code: PAGZOO, LCD (Å): 19.55554
      There are more rows remaining, say 'give me another X' to see more.
```

Figure S41: Structure Q&A of MOFh6: Finding structures within the numerical range

```
Query> Compare the surface areas of VUJBEI and QOWTIG
INFO - Query analysis: {'query_type': 'comparison', 'materials': ['VUJBEI', 'QOWTIG'], 'properties':
['surface area'], 'operation': None, 'range': None, 'uses_context': True, 'reasoning': [], 'page_size':
None}
Answer> The surface areas for the materials are as follows:
        \- **VUJBEI**:
         \- Accessible Surface Area (m²/cm³): 2003.25
         \- Accessible Surface Area (m²/g): 3366.56
        \- **QOWTIG**:
         \- Accessible Surface Area (m²/cm³): 1474.01
         \- Accessible Surface Area (m²/g): 1538.65
        **Comparison**: VUJBEI has a higher accessible surface area than QOWTIG in both m²/cm³ and m²/g
measurements.

Query> who is the boss between SALFAO and SAZCED
INFO - Query analysis: {'query_type': 'comparison', 'materials': ['SALFAO', 'SAZCED'], 'properties': [],
'operation': None, 'range': None, 'uses_context': True, 'reasoning': ['User is asking for a comparison
between two materials, SALFAO and SAZCED.'], 'page_size': None}
Answer> Here is a brief comparison between SALFAO and SAZCED:
        1. **SALFAO**
           - CCDC_number: 829986
           - Chemical_name: catena-[tetrakis(μ3-Iodo)-tris(μ3-3-(4-methylphenyl)-5-(4-pyridyl)pyrazolato-
N,N1,N2)-ammine-hepta-copper(i)]
           - Density: 2.3638 g/cm³
           - Accessible_Surface_Area: 0.0 m²/g
        2. **SAZCED**
           - CCDC_number: 867556
           - Chemical_name: catena-(tetrakis(μ2-thiocyanato)-tetrakis(isonicotinic acid)-di-nickel
pyrene)
           - Density: 0.66533 g/cm³
           - Accessible_Surface_Area: 3362.1 m²/g
        Based on the properties, SAZCED has a significantly higher accessible surface area and lower
density compared to SALFAO. If you need more details or additional rows, please let me know!
```

Figure S42: Structure Q&A of MOFh6: Comparing properties between structures

```
Query> What is the average density?
INFO - Query analysis: {'query_type': 'statistical', 'materials': [], 'properties': ['density'],
'operation': {'type': 'mean', 'value': None}, 'range': {'min': None, 'max': None}, 'uses_context': False,
'reasoning': [], 'page_size': None}
Answer> The average density is 1.707 g/cm³.

Query> Show material with highest gravimetric surface area
INFO - Query analysis: {'query_type': 'comparison', 'materials': [], 'properties': ['Gravimetric Surface
Area'], 'operation': {'type': 'max', 'value': None}, 'range': {'min': None, 'max': None}, 'uses_context':
True, 'reasoning': ['User is seeking a material with the highest gravimetric surface area.'],
'page_size': None}
Answer> The material with the highest gravimetric surface area is:
        \- **CCDC Code**: YAKLON
        \- **Accessible Surface Area**: 9556.14 m²/g
        If you need more information or additional materials, feel free to ask!
```

Figure S43: Structure Q&A of MOFh6: Numerical statistics of the dataset

```
Query> Can you find materials with PLD of 7.5-10 Å, LCD of 10-16 Å, VSA of 2000-2400 m2 cm-3
INFO - Query analysis: {'query_type': 'range', 'materials': [], 'properties': [], 'operation': {'type':
None, 'value': None}, 'range': {'min': {'PLD Å': 7.5, 'LCD Å': 10, 'VSA m2/cm3': 2000}, 'max': {'PLD Å':
10, 'LCD Å': 16, 'VSA m2/cm3': 2400}}, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> I found a material that matches your criteria:
        - **CCDC Code**: VUJBEI
        If you would like more details or additional materials, please let me know!

Query> is any others?
- INFO - Query analysis: {'query_type': 'chat', 'materials': [], 'properties': [], 'operation': {'type':
None, 'value': None}, 'range': {'min': {'PLD': 7.5, 'LCD': 10, 'VSA': 2000}, 'max': {'PLD': 10, 'LCD':
16, 'VSA': 2400}}, 'uses_context': True, 'reasoning': [], 'page_size': None}
Answer> Here are some materials that fit your criteria:
        1. **CCDC_code**: ACUCAC
           - **Chemical_name**: catena-(octakis(μ2-3-(2-(4-pyridyl)ethenyl)benzoato)-tetra-zinc) (3-(2-
(4-pyridyl)ethenyl)benzoic acid monohydrate)
           - **PLD (Å)**: 9.19678
           - **LCD (Å)**: 11.76754
           - **Density (g/cm³)**: 0.321479
           - **Accessible_Surface_Area (m²/cm³)**: 2221.1
        2. **CCDC_code**: ADATUW
           - **Chemical_name**: catena-((μ3-tris(4-(1-(2-pyridin-2-ylhydrazono)ethyl)-phenyl)amine)-
silver tetrafluoroborate acetonitrile dichloromethane solvate monohydrate)
           - **PLD (Å)**: 8.86044
           - **LCD (Å)**: 10.34629
           - **Density (g/cm³)**: 0.640612
           - **Accessible_Surface_Area (m²/cm³)**: 2122.36
        3. **CCDC_code**: ADAVAE
           - **Chemical_name**: catena-((μ3-tris(4-(1-(2-pyridin-2-ylhydrazono)ethyl)-phenyl)amine)-
silver tetrafluoroborate dichloromethane N,N-dimethylformamide solvate monohydrate)
           - **PLD (Å)**: 8.80719
           - **LCD (Å)**: 10.21446
           - **Density (g/cm³)**: 0.655338
           - **Accessible_Surface_Area (m²/cm³)**: 2079.91
        4. **CCDC_code**: ADAVEI
           - **Chemical_name**: catena-[tris(μ3-tris(4-(1-(2-pyridin-2-ylhydrazono)ethyl)phenyl)amine)-
tri-silver tris(tetrafluoroborate) acetonitrile chloroform methanol solvate monohydrate]
           - **PLD (Å)**: 8.70819
           - **LCD (Å)**: 10.15021
           - **Density (g/cm³)**: 0.656268
           - **Accessible_Surface_Area (m²/cm³)**: 2127.06
        5. **CCDC_code**: AGOVID
           - **Chemical_name**: catena-(tris(μ4-Biphenyl-4,4'-dicarboxylato)-hexakis(μ3-4-(pyridin-4-
yl)benzoato)-bis(μ3-oxo)-hexa-indium dinitrate)
           - **PLD (Å)**: 8.64841
           - **LCD (Å)**: 12.48057
           - **Density (g/cm³)**: 0.64221
           - **Accessible_Surface_Area (m²/cm³)**: 2023.25
        If you need more information or additional entries, please let me know!
```

Figure S44: Structure Q&A of MOFh6: Comprehensive structural query based on multiple structural characteristics

```
**English**
Query> What is the PLD of VUJBEI?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': False, 'reasoning': [],
'page_size': None}
Answer> The PLD of VUJBEI is 4.92727 Å.

**German**
Query> Was ist die PLD von VUJBEI?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> Die PLD von VUJBEI beträgt 4.92727 Å.

**Chinese**
Query> VUJBEI的PLD是多少?
Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD'], 'operation':
{'type': None, 'value': None}, 'range': {}, 'uses_context': False, 'reasoning': [], 'page_size': None}
Answer> VUJBEI的PLD是4.92727 Å。

**Russian**
Query> Сколько PLD у VUJBEI?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> PLD у Vujbei составляет 4.92727 Å.

**French**
Query> Quel est le PLD de VUJBEI?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> Le PLD de VUJBEI est de 4.92727 Å.

**Japanese**
Query> VUJBEIのPLDはいくらですか?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> VUJBEIのPLDは4.92727 Åです。

**Korean**
Query> VUJBEI의 PLD는 얼마입니까?
INFO - Query analysis: {'query_type': 'property', 'materials': ['VUJBEI'], 'properties': ['PLD (Å)'],
'operation': {'type': None, 'value': None}, 'range': {}, 'uses_context': True, 'reasoning': [],
'page_size': None}
Answer> VUJBEI의 PLD는 4.92727 Å입니다.
```

Figure S45: Multi language support of MOFh6

Field Name	Data Type	Description	Example Values
query_type	String	Query type identifier	"property", "comparison", "statistical", "range", "chat", "greeting", "reset", "unknown", "paging"
materials	Array	Material identifiers mentioned in query	["VUJBEI", "QOWTIG"], ["MOF-5"]
properties	Array	Material properties mentioned in query	["PLD (Å)"], ["density", "surface area"]
operation	Object	Statistical or comparison operation info	{"type": "mean", "value": null}, {"type": "max", "value": null}
operation.type	String	Operation type	"mean", "max", "min", null
operation.value	Number/null	Value related to operation	null (in most cases), may contain specific values
range	Object	Constraints for range queries	{"min": {"PLD": 7.5, "LCD": 10}, "max": {"PLD": 10, "LCD": 16}}
range.min	Object	Minimum value constraints	{"PLD": 7.5, "LCD": 10, "VSA": 2000}
range.max	Object	Maximum value constraints	{"PLD": 10, "LCD": 16, "VSA": 2400}
uses_context	Boolean	Whether context is used	true, false
reasoning	Array	Reasoning process for query parsing	["User is asking for a specific property of the material mentioned in the last query"]
page_size	Integer/null	Page size for pagination queries	5, 10, null
last_query	String	Previous query content	"What is the PLD of VUJBEI?"
last_materials	Array	Materials involved in the last query	["VUJBEI"]
last_properties	Array	Properties involved in the last query	["PLD (Å)"]
last_result	Object	Result set of the last query	DataFrame object
paged_index	Integer	Current index for pagination query	0, 5, 10
query_history	Array	Query history records	[{"question": "...", "analysis": {...}, "materials": [...], "properties": [...]}]

Figure S46: MOFh6 system query parameter fields

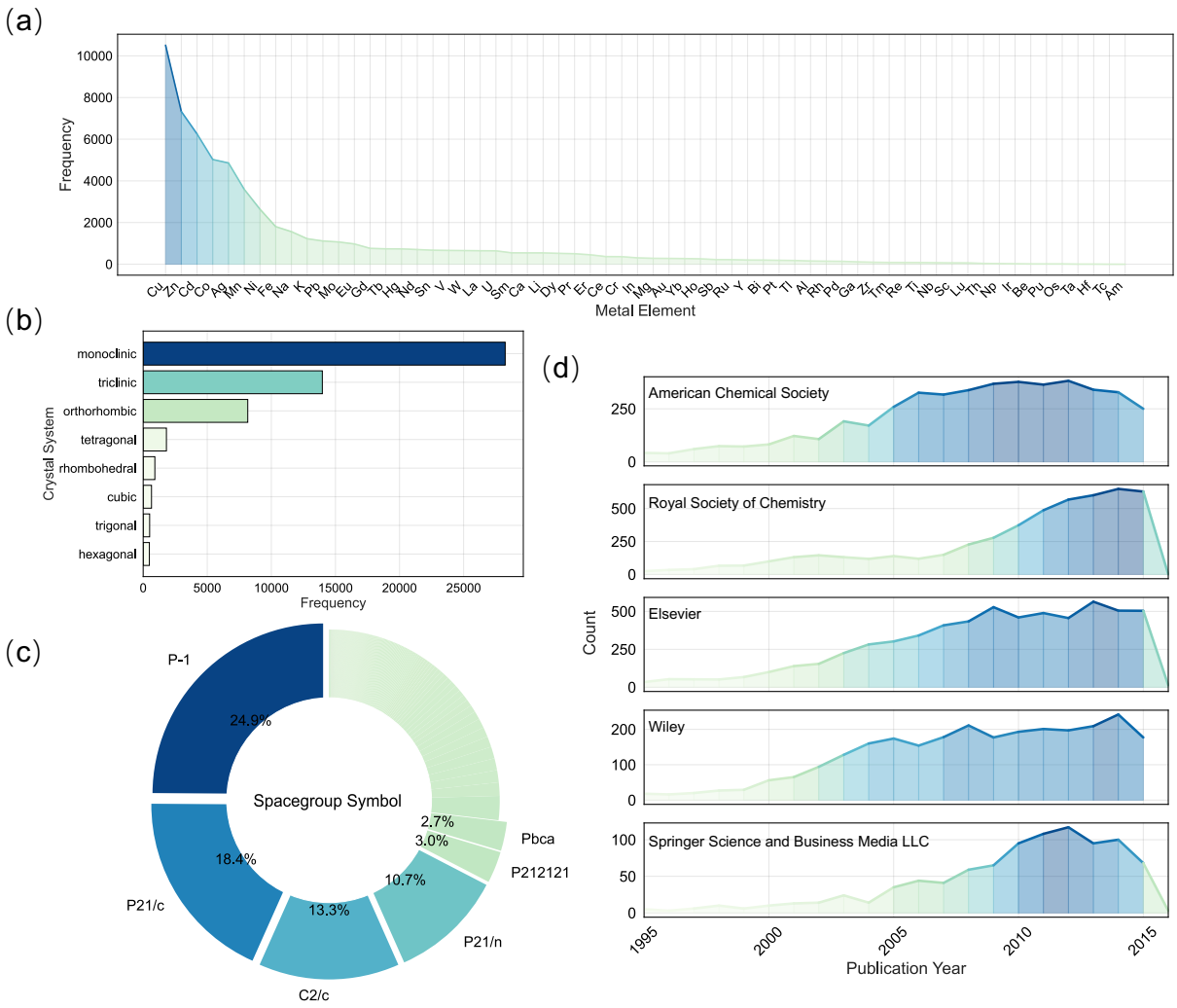

Figure S47: Data mining statistics, frequency of occurrence of metal elements that make up MOFs (a), crystal system (b), space group symbol (c), publisher publication volume (d)

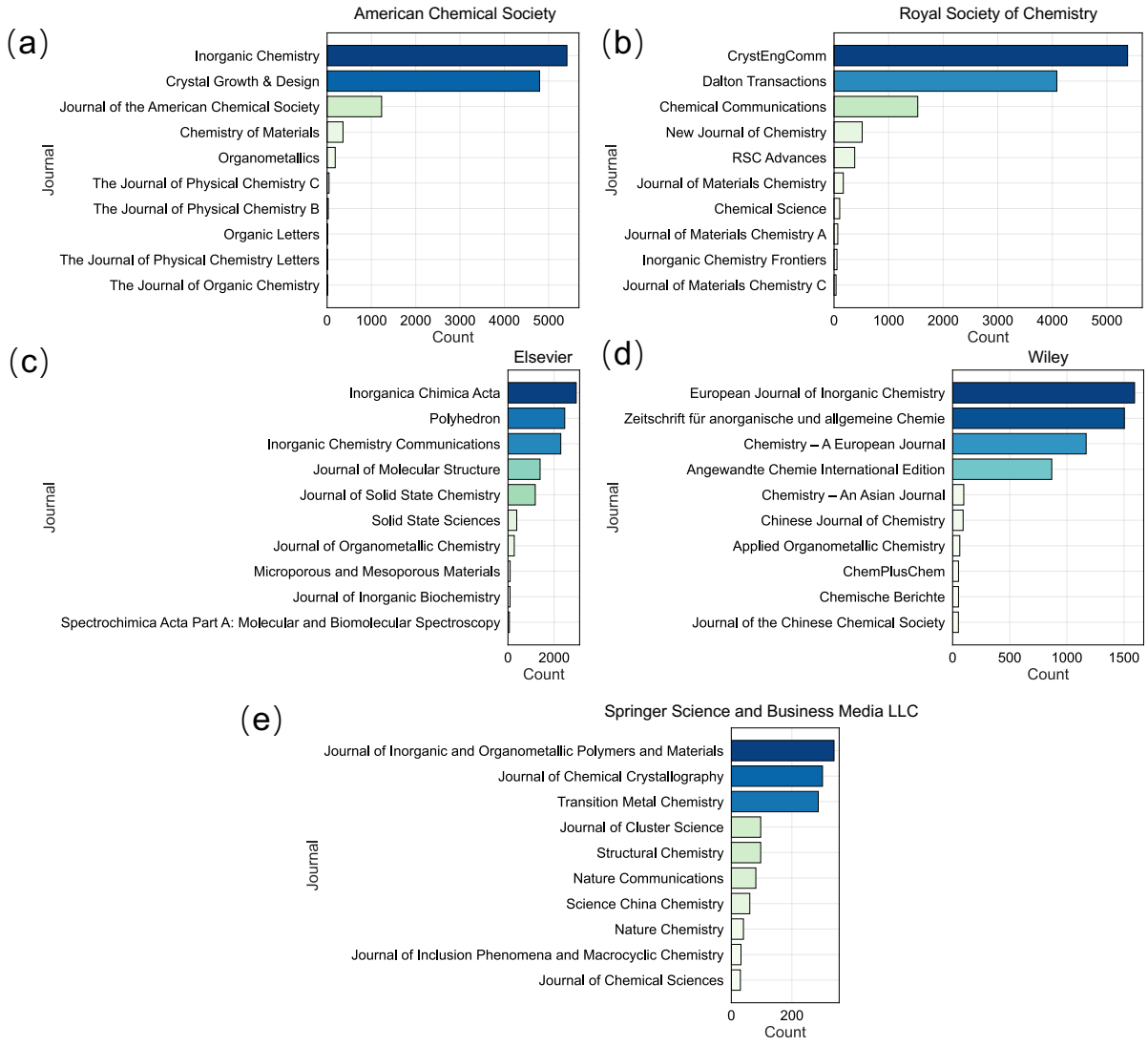

Figure S48: Statistics of the top ten journals from different publishers, ACS (a), RSC (b), Elsevier (c), Wiley (d), Springer (e)

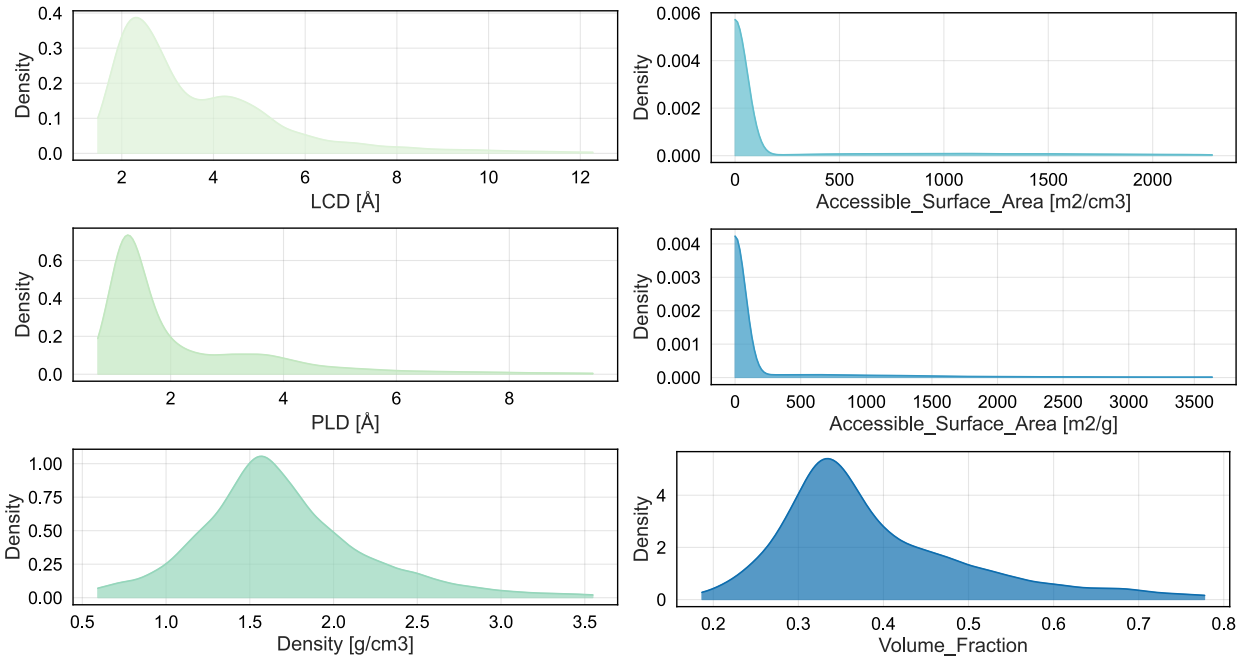

Figure S49: Structural properties of MOFs calculated by *Zeo++*

Welcome to MOF Analysis System!

Available Commands:

'How to synthesize <code/name>' - Obtain the synthesized text

'process pdf <path>' - Process a PDF document

'workflow <path>' - Run analysis workflow on processed PDF

'show saved' - Show saved documents

'download cif <code>' - Download CIF file for a CCDC code

'visualize <code>' - Visualize crystal structure for a CCDC code

'q!' - Exit the system

download cif BAZBOU02

Successfully downloaded from Willlzh/ccdccif1

Successfully downloaded CIF file for BAZBOU02

To visualize the structure, type:

visualize BAZBOU02

visualize BAZBOU02

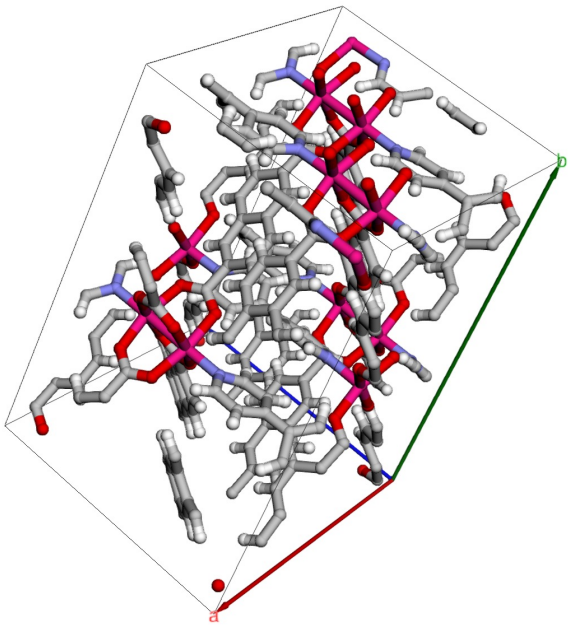

Figure S50: CIF download and visualization of MOFh6

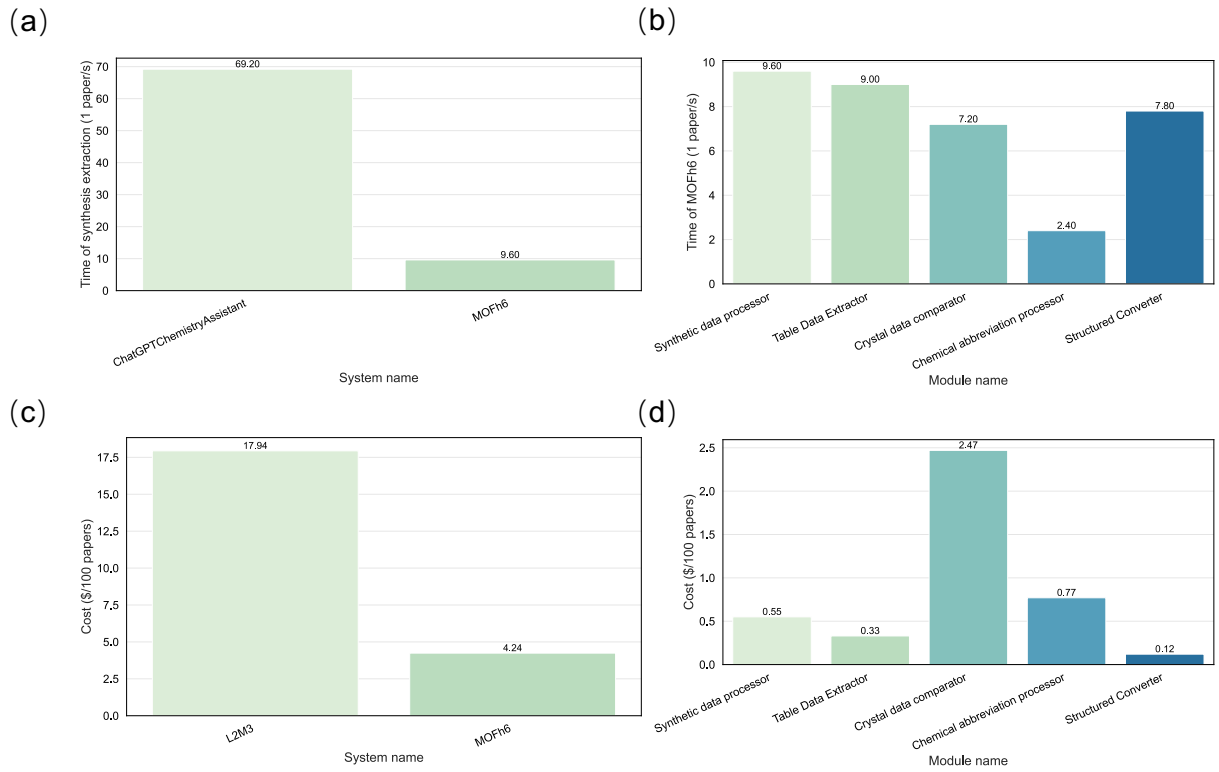

Figure S51: Cost comparison between MOFh6 and similar works

Figure S52: MOFh6 data mining processing final localization for specified MOFs synthesis requests

Table S1: The core agents of MOFh6

Agent name	Core model algorithm	Output format	System level
Synthetic data parsing agent	GPT-4o-mini fine-tuning	.xlsx	1
Table data parsing agent	GPT-4o-mini prompt + Structured parsing	tables_*.json	
Crystal data comparison agent	GPT-4o prompt + Structural similarity calculation	comparison_*.json	
Chemical abbreviation resolution agent	GPT-4o prompt + Regular expression pattern matching	acronym_results_*.json	2
Post processor	GPT-4o mini prompt + Text segmentation + File organization management	identifier_*.txt	
Result Generator agent	GPT-4o-mini prompt + BM25 algorithm	final_output_*.txt	3
Structured conversion agent	GPT-4o-mini prompt + Markdown conversion	structure_*.md	

Table S2: Models for evaluating the performance of MOFh6

Level	Index	Model
Sentence	Description of synthetic paragraphs	BiomedBERT
Data frame	Metal Source	PubMedBERT
	Organic Linkers Source	
	Modulator Source	
	Solvent Source	
	Quantity of Metal	
	Quantity of Organic Linkers	
	Quantity of Modulator	
	Quantity of Solvent	
	Synthesis Temperature	all-mpnet-base-v2
	Synthesis Time	
	Crystal Morphology	
	Yield	
	Equipment	