HHS Public Access

Author manuscript

J Am Chem Soc. Author manuscript; available in PMC 2024 May 06.

Published in final edited form as:

J Am Chem Soc. 2023 August 16; 145(32): 18048–18062. doi:10.1021/jacs.3c05819.

ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis

Zhiling Zheng,
Department of Chemistry, University of California, Berkeley, California 94720, United States

Kavli Energy Nanoscience Institute and Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States

Oufan Zhang,
Department of Chemistry, University of California, Berkeley, California 94720, United States

Kavli Energy Nanoscience Institute, Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, Department of Electrical Engineering and Computer Sciences, Department of Mathematics, Department of Statistics, and School of Information, University of California, Berkeley, California 94720, United States;

KACST–UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

Christian Borgs,
Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, United States

Jennifer T. Chayes,
Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, Department of Electrical Engineering and Computer Sciences, Department of Mathematics, Department of Statistics, and School of Information, University of California, Berkeley, California 94720, United States

Omar M. Yaghi
Department of Chemistry, University of California, Berkeley, California 94720, United States

Corresponding Author Omar M. Yaghi – *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute and Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States; KACST–UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*; yaghi@berkeley.edu.

The authors declare no competing financial interest.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c05819.

Detailed instructions and design principles for ChemPrompt Engineering and the specifics of the prompts employed in the ChatGPT Chemistry Assistant for text mining and other chemistry-related tasks; additional information on the ChatGPT-assisted coding and data processing methods; and an extensive explanation of the machine learning models and methods used as well as the steps involved in setting up the MOF chatbot based on the MOF synthesis condition data set (PDF)

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c05819

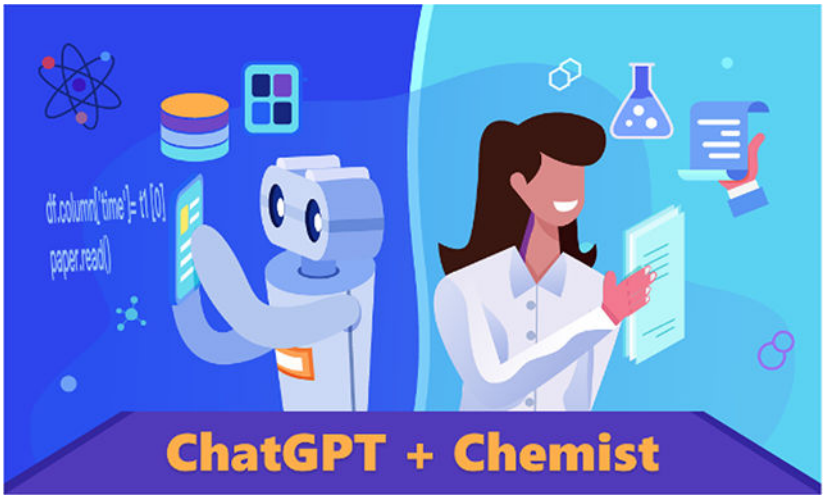
Kavli Energy Nanoscience Institute and Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States

KACST–UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia

Abstract

We use prompt engineering to guide ChatGPT in the automation of text mining of metal–organic framework (MOF) synthesis conditions from diverse formats and styles of the scientific literature. This effectively mitigates ChatGPT's tendency to hallucinate information, an issue that previously made the use of large language models (LLMs) in scientific fields challenging. Our approach involves the development of a workflow implementing three different processes for text mining, programmed by ChatGPT itself. All of them enable parsing, searching, filtering, classification, summarization, and data unification with different trade-offs among labor, speed, and accuracy. We deploy this system to extract 26 257 distinct synthesis parameters pertaining to approximately 800 MOFs sourced from peer-reviewed research articles. This process incorporates our ChemPrompt Engineering strategy to instruct ChatGPT in text mining, resulting in impressive precision, recall, and F1 scores of 90–99%. Furthermore, with the data set built by text mining, we constructed a machine-learning model with over 87% accuracy in predicting MOF experimental crystallization outcomes and preliminarily identifying important factors in MOF crystallization. We also developed a reliable data-grounded MOF chatbot to answer questions about chemical reactions and synthesis procedures. Given that the process of using ChatGPT reliably mines and tabulates diverse MOF synthesis information in a unified format while using only narrative language requiring no coding expertise, we anticipate that our ChatGPT Chemistry Assistant will be very useful across various other chemistry subdisciplines.

Graphical Abstract



INTRODUCTION

The dream of chemists is to create matter in the hope of advancing human knowledge for the betterment of society.[1,2] As we stand on the precipice of the age of artificial general intelligence (AGI), the potential for synergy between AI and chemistry is vast and promising.[3,4] The idea of creating AI-powered chemistry assistants offers unprecedented opportunities to revolutionize the landscape of chemistry research by applying knowledge across various disciplines, efficiently processing labor-intensive and time-consuming tasks such as literature searches, compound screening, and data analysis. AI-powered chemistry may ultimately transcend the limits of human cognition.[5–8]

Identifying chemical information for compounds, including ideal synthesis conditions and physical and chemical properties, has been a critical endeavor in chemistry research. The comprehensive summary of chemical information from literature reports, such as publications and patents, and their subsequent storage in an organized database format is the next logical and necessary step toward the discovery of materials.[9] The challenge lies in efficiently mining the vast amount of available literature to obtain valuable information and insights. Traditionally, specialized natural language processing (NLP) models have been employed to address this issue.[10–14] However, these approaches can be labor-intensive and necessitate expertise in coding, computer science, and data science. Furthermore, they are less generalizable, requiring the program to be rewritten when the target changes. The advent of large language models (LLMs), such as GPT-3, GPT-3.5, and GPT-4, has the potential to fundamentally transform this process and revolutionize the routine of chemistry research in the next decade.[9,15–18]

Herein, we demonstrate that LLMs, including ChatGPT based on the GPT-3.5 and GPT-4 models, can act as chemistry assistants to collaborate with human researchers, facilitating text mining and data analysis to accelerate the research process. To harness the power of what we termed the ChatGPT Chemistry Assistant (CCA), we provide a comprehensive guide on ChatGPT prompt engineering for chemistry-related tasks, making it accessible to researchers regardless of their familiarity with machine learning, thus bridging the gap between chemists and computer scientists. In this report, we present (1) a novel approach to using ChatGPT for text mining the synthesis conditions of metal–organic frameworks (MOFs), which can be easily generalizable to other contexts requiring minimal coding knowledge and operating primarily on verbal instructions; (2) an assessment of ChatGPT's intelligence in literature text mining through accuracy evaluation and its ability for data refinement; and (3) utilization of the chemical synthesis reaction data set obtained from text mining to train a model capable of predicting reaction results as crystalline powder or single crystals. Furthermore, we demonstrate that the CCA chatbot can be tuned to specialize in answering questions related to MOF synthesis based on literature conditions, with minimal hallucinations. This study underscores the transformative potential of ChatGPT and other LLMs in the realm of chemistry research, offering new avenues for collaboration and accelerating scientific discovery.

MATERIALS AND METHODS

Design Considerations for ChatGPT-Based Text Mining.

In curating research papers for ChatGPT to read and extract information, it is imperative to account for the diversity in MOF synthesis conditions, such as variations in metal sources, linkers, solvents, and equipment as well as the different writing styles employed. Notably, the absence of a standardized format for reporting MOF synthesis conditions leads to variable reporting templates by research groups and journals. Indeed, by incorporating a broad spectrum of narrative styles, we can examine ChatGPT's robustness in processing information from heterogeneous sources. On the other hand, it is essential to recognize that the challenge of establishing unambiguous criteria to identify MOF compounds in the literature may lead to the inadvertent inclusion of some non-MOF compounds reported in earlier publications that are nonporous inorganic complexes and amorphous coordination polymers (included in some MOF data sets). As such, maintaining a balance between quality and quantity is vital, and prioritizing the selection of high-quality and well-cited papers, rather than incorporating all associated papers indiscriminately, can ensure that the text mining of MOF synthesis conditions yields reliable and accurate data.

Moreover, papers discussing postsynthetic modifications, catalytic reactions of MOFs, and MOF composites are not directly pertinent to our objective of identifying MOF synthesis conditions. Hence, such papers have been excluded. Another consideration is that MOFs can be synthesized as both microcrystalline powders and single crystals, both of which should be regarded as valid candidates for our data set. Utilizing the above-mentioned selection criteria, we narrowed our selection to 228 papers from an extensive pool of MOF papers, retrieved from the Web of Science, the Cambridge Structure Database MOF subset,[19] and the CoreMOF database.[20,21] This sample represents a diverse range of MOF synthesis conditions and narrative styles.

To enable ChatGPT to process each paper, we devised three different approaches analogous to human paper reading: (1) locating potential sections containing synthesis conditions within the document, (2) confirming the presence of synthesis conditions in the identified sections, and (3) extracting synthesis parameters one by one. For our ChatGPT Chemistry Assistant, these steps are accomplished through filtering, classification, and summarization (Figure 1).

In Process 1, we developed prompts to guide ChatGPT in summarizing text from designated experimental sections contained in those papers. To replace the need for human intervention to obtain synthesis sections, in Process 2, we designed a method for ChatGPT to categorize text inputs as either "experimental section" or "nonexperimental section", enabling it to generate experimental sections for summarization. In Process 3, we further devised a technique to swiftly eliminate irrelevant paper sections, such as references, titles, and acknowledgments, which are unlikely to encompass comprehensive synthesis conditions. This accelerates the processing speed for the later classification task. As such, in Process 1, ChatGPT is solely responsible for summarizing and tabulating synthesis conditions and requires one or more paragraphs of experimental text as input, while Processes 2 and 3 can be considered to be an "automated paper reading system". While Process 2 entails

a thorough examination of the entire paper to scrutinize each section, the more efficient Process 3 rapidly scans the entire paper, removing the least relevant portions and thereby reducing the number of paragraphs that ChatGPT must meticulously analyze.

Prompt Engineering.

In the realm of chemistry-related tasks, ChatGPT's performance can be significantly enhanced by employing prompt engineering (PE)—a meticulous approach to designing prompts that steer ChatGPT toward generating precise and pertinent information. We propose three fundamental principles in prompt engineering for chemistry-focused applications, denoted as ChemPrompt Engineering:

1. *Minimizing Hallucination*, which entails the formulation of prompts to avoid eliciting fabricated or misleading content from ChatGPT. This is particularly important in the field of chemistry, where the accuracy of information can have significant implications on research outcomes and safety. For instance, when asked to provide synthesis conditions for MOFs without any additional prompt or context, ChatGPT may recognize that MOF-99999 does not exist but will generate fabricated conditions for existing compounds with names such as MOF-41, MOF-419, and MOF-519. We should note that with additional prompts followed after the question, it is possible to minimize hallucination and force ChatGPT to answer the questions based on its knowledge (Tables 1 and 2). Furthermore, we demonstrate that with well-designed prompts and context, hallucination occurrences can be minimized (Supporting Information, Section S2.1). We note that this should be the first and foremost principle to follow when designing prompts for ChatGPT to perform in handling text and questions relevant to chemical information.

2. *Implementing Detailed Instructions*, whereby explicit directions are provided in the prompt to assist ChatGPT in understanding the context and desired response format. By incorporating detailed guidance and context into the prompts, we can facilitate a more focused and accurate response from ChatGPT. In chemistry-related tasks, this approach narrows down the potential answer space and reduces the likelihood of irrelevant or ambiguous responses. For example, we can specify not to include any organic linker synthesis conditions and focus solely on MOF synthesis (Supporting Information, Figure S8). In this case, we found that ChatGPT can recognize the features of organic linker synthesis and differentiate them from MOF synthesis. With proper prompts, information from organic linker synthesis will not be included. Additionally, instructions can provide step-by-step guidance, which has proven effective when multiple tasks are included in one prompt (Supporting Information, Section S2.2).

3. *Requesting Structured Output*, which includes the incorporation of an organized and well-defined response template or instructions to facilitate data extraction. We emphasize that this principle is particularly valuable in the context of chemistry, where data can often be complex and multifaceted. Structured output enables the efficient extraction and interpretation of critical information, which in turn can significantly contribute to the advancement of research and

knowledge in the field. Taking synthesis condition extraction as an example, without clear instructions on the formatted output, ChatGPT can generate a table, list-like bullet points, or a paragraph, with the order of parameters such as the reaction temperature, reaction time, and solvent volume not being uniform, making it challenging for later sorting and storage of the data. This can be easily improved by explicitly asking it to generate a table and providing a fixed header to start with a prompt (Supporting Information, Section S2.3). By incorporating these principles, the resulting prompt can ensure that ChatGPT yields accurate and reliable results, ultimately enhancing its utility in tackling complex chemistry-related tasks (Figure 2). We further employ the idea of interactive prompt refinement, in which we start by asking ChatGPT to write a prompt to instruct itself by giving it preliminary descriptions and information (Supporting Information, Figure S15). Through conversation, we add more specific details and considerations to the prompt, testing it with some texts, and once we obtain output, we provide feedback to ChatGPT and ask it to improve the quality of the prompt (Supporting Information, Section S2.4).

As there has been almost no literature systematically discussing prompt engineering in chemistry and the fact that this field is relatively new, we provide a comprehensive step-by-step ChemPrompt Engineering guide for beginners to start with, including numerous chemistry-related examples in the Supporting Information, Section S2. At present, everyone is at the same starting point, and no one possesses exclusive expertise in this area. It is our hope that this work will stimulate the development of more powerful prompt engineering skills and help every chemist quickly understand the art of ChemPrompt Engineering, thereby advancing the field of chemistry at large.

Process 1: Synthesis Conditions Summarization.

One revolutionary aspect of ChatGPT is its specialized domain knowledge due to its extensive pretrained text corpus, which enables an understanding of chemical nomenclature and reaction conditions.[18] In contrast to traditional NLP methods, ChatGPT requires no additional training for named entity recognition and can readily identify inorganic metal sources, organic linkers, solvents, and other compounds within a given experimental text. Another notable feature is ChatGPT's ability to recognize and associate compound abbreviations (e.g., DMF) with their full names (*N*,*N*-dimethylformamide) within the context of MOF synthesis (Supporting Information, Figure S5). This capability is crucial as the use of different abbreviations for the same compound can inflate the number of "unique compounds" in the data set after text mining, leading to redundancy without providing new information. This challenge is difficult to address using traditional NLP methods or packages as no model can inherently discern that DMF and *N*,*N*-dimethylformamide are the same compound without a manually curated dictionary of chemical abbreviations. Although ChatGPT may not cover all abbreviations, its proficiency in identifying and associating the most common ones, such as DEF, DI water, EtOH, and CH_3CN with their full names, enhances data consistency and reduces redundancy. This, in turn, facilitates data retrieval and analysis, ensuring that different names of the same compound are treated as a single entity with its unique chemical identity and information.

Our first goal is to develop a ChatGPT-based AI assistant that demonstrates high performance in converting a given experimental section paragraph into a table containing all synthesis parameters (Supporting Information, Figure S22). To design the prompt for this purpose, we incorporate the three principles discussed earlier into ChemPrompt Engineering (Figure 2). The rationale for using tabulation as the output for synthesis condition summarization is that the tabular format simplifies subsequent data sorting, analysis, and storage. In terms of the choice of 11 synthesis parameters, we include those deemed most important and non-negligible for each MOF synthesis. Specifically, these parameters encompass metal sources and quantities, dictating metal centers in the framework and their relative concentrations; the linker and its quantity, which affect connectivity and pore size within the MOF; the modulator and its quantity or volume, which can fine-tune the MOF's structure by impacting the nucleation and growth of the MOF in the reaction; the solvent and its volume, which can influence both the crystallization process and the final MOF structure; and the reaction temperature and duration, which are vital parameters governing the kinetics and thermodynamics of MOF formation in each synthesis. In our prompt, we also account for the fact that some papers may report multiple synthesis conditions for the same compound and instruct ChatGPT to use multiple rows to include each variation. For multiple units of the same synthesis parameters, such as when molarity mass and weight mass are both reported, we encourage ChatGPT to include them in the same cell, separated by a comma, which can be later streamlined depending on the need. If any information is not provided in the sections, e.g., most MOF reactions may not involve the use of modulators and some papers may not specify the reaction time, then we expect ChatGPT to answer "N/A" for that parameter. Importantly, to eliminate non-MOF synthesis conditions such as organic linker synthesis, postsynthetic modification, and catalysis reactions, which are not helpful for studying MOF synthesis reactions, we simply add one line of narrative instruction, asking ChatGPT to ignore these types of reactions and focus solely on MOF synthesis parameters. Notably, this natural language-based instruction is highly convenient, requiring no complex and laborious rule-based code to identify unwanted cases and filter them out, and is friendly to researchers without coding experience.

The finalized prompts for Process 1 consist of three parts: (i) a request for ChatGPT to summarize and tabulate the reaction conditions and use only the text or information provided by humans, which adheres to Principle 1 to minimize hallucination; (ii) a specification of the output table's structure, enumerating expectations and handling instructions, which follows Principles 2 and 3 for detailed instructions and structured output requests; and (iii) the context, consisting of MOF synthesis reaction condition paragraphs from experimental sections or the Supporting Information in research articles. Note that parts (i) and (ii) are fixed prompts, while part (iii) is considered to be "input". The combined prompt results in a single question-and-answer interaction, allowing ChatGPT to generate a summarization of the given synthesis conditions as output.

Process 2: Synthesis Paragraph Classification.

The next question to be answered is, "if ChatGPT is given an entire research article, can it correctly locate the experimental sections?" The objective of Process 2 is to accept an entire research paper as input and selectively forward paragraphs containing

chemical experiment details to the next assistant for summarization. However, locating the experimental synthesis section within a research paper is a complex task, as simple techniques, such as keyword searches, often prove insufficient. For instance, the synthesis of MOFs may be embedded within the Supporting Information or combined with organic linker synthesis. In earlier publications, the synthesis information might appear as a footnote. Furthermore, different journals or research groups utilize varying section titles, including “Experimental”, “Methods”, “General Methods and Materials”, “Experimental Methods”, “Synthesis and Characterization”, “Synthetic Procedures”, “Methods Summary”, and more. Manually enumerating each case is labor-intensive, especially when synthesis paragraphs may be dispersed with non-MOF synthesis conditions or instrument details. Even a human might take considerable time to identify the correct section.

To address this challenge and enable ChatGPT to accurately discern synthesis details within a lengthy research paper, we draw inspiration from the human process. A chemistry Ph.D. student, when asked to locate the MOF synthesis section in a new research paper, would typically start with the first paragraph and ask themselves if it contains synthesis parameters. They would then draw upon prior knowledge from previously read papers to determine whether the section is experimental. This process is repeated paragraph by paragraph until the end of the Supporting Information is reached, with no guarantee that additional synthesis details will not be encountered later. To train ChatGPT similarly, we prompt it to read paper sections incrementally, focusing on one or two paragraphs at a time. Using a few-shot prompt strategy, we provided ChatGPT with a couple of example cases of both synthesis and nonsynthesis paragraphs and asked it to classify the sections it read as either “Yes” (synthesis paragraph) or “No” (nonsynthesis paragraph). The ChatGPT Chemistry Assistant would then continue processing the research paper section by section, passing only the paragraphs labeled as “Yes” to the following assistant for summarization.

This few-shot prompt strategy is more convenient than traditional approaches, which require researchers to manually identify and label a large number of paragraphs as “Synthesis Paragraphs” and train their models accordingly. In fact, ChatGPT can even perform such classification using a zero-shot prompt strategy with detailed descriptions of what a “Synthesis Paragraph” should look like and contain. However, we have found that providing four or five short examples in a few-shot prompt strategy enables ChatGPT to identify the features of synthesis paragraphs more effectively, streamlining the classification process (Supporting Information, Figure S24).

The finalized prompt for Process 2 comprises three parts: (i) a request for ChatGPT to determine whether the provided context includes a comprehensive MOF synthesis, answering only with “Yes” or “No”; (ii) some example contexts labeled as “Yes” and others labeled as “No”; and (iii) the context to be classified, consisting of one or more research article paragraphs. Similar to Process 1’s prompt, parts (i) and (ii) are fixed, while part (iii) is replaced with independent sections from the paper to be classified. The entire research article is parsed into sections of 100–500 words, which are iteratively incorporated into the prompt and sent separately to ChatGPT for a “Yes” or “No” response. Each prompt represents a one-time conversation, and ChatGPT cannot view answers from previous prompts, preventing potential bias in its decision making for the current prompt.

Process 3: Text Embeddings for Search and Filtering.

Text embeddings are high-dimensional vector representations of text that capture semantic information, enabling quantification of the relatedness of textual content.[22,23] The distance between these vectors in the embedded space correlates with the semantic similarity between corresponding text strings, with smaller distances indicating greater relatedness.[24,25] While Process 2 can automatically read and summarize papers, it must evaluate every section to identify synthesis paragraphs. To expedite this process, we developed Process 3, which filters sections least likely to contain synthesis parameters using OpenAI embeddings before exposing the article to the classification assistant in Process 2. To achieve this, we employed a two-step approach to construct Process 3: first, parsing all papers and converting each segment into embeddings and second, calculating and ranking the similarity scores of each segment based on their relevance to a predefined prompt encapsulating the synthesis parameter.

In particular, we partitioned the 228 research articles into 18 248 individual text segments (Supporting Information, Figures S30–S32). Each segment was converted into 1536-dimensional text embedding using OpenAI's *text-embedding-ada-002*, a simple but efficient model for this process (Supporting Information, Figures S33–S35). These vectors were stored for future use. To identify segments most likely to contain synthesis parameters, we employed an interactive prompt refinement strategy (Supporting Information, Section S2.4), consulting with ChatGPT to optimize the prompt. The prompt used in Process 3, unlike previous prompts, served as a text segment for search and similarity comparison rather than instructing ChatGPT (Supporting Information, Figure S25). Next, the embeddings of all 18 248 text segments were compared with the prompt's embedding, and a relevance score was assigned to each segment based on the cosine similarity between the two embeddings. Highly relevant segments were passed on to a classification assistant for further processing, while low-similarity segments were filtered out (Figure 1).

To evaluate the effectiveness of this approach, we conducted a visual exploration of our embedding data (Figure 3). By reducing the vectors' dimensionality, we observed distinct clusters corresponding to different topics. Notably, we identified distinct clusters related to topics such as gas sorption, literature references, characterization, structural analysis, and crystallographic data, which were separate from the synthesis cluster. This observation strongly supports the efficiency of our embedding-based filtering strategy. However, this strategy, while effective at filtering out less relevant text and passing segments of mid- to high relevance to the subsequent classification assistant, cannot directly search for synthesis paragraphs to feed to the summarization assistant, thus bypassing the classification assistant. In other words, the searching-to-classifying-to-summarizing pipeline cannot be simplified to a searching-to-summarizing pathway due to the inherent search limitations of the embeddings. As shown in Figure 3, embeddings alone may not accurately identify all relevant synthesis sections, particularly when they contain additional information, such as characterization and sorption data. The presence of these elements in a synthesis section can reduce its similarity score and its proximity to the center of the synthesis cluster. Points between the synthesis and characterization or crystallographic data clusters may not have the highest similarity scores and could be missed. However, by filtering only the lowest scores,

midrelevance points are retained and passed to the classification assistant, which can more accurately classify ambiguous content.

ChatGPT-Assisted Python Code Generation and Data Processing.

Rather than relying on singular, time-consuming conversations with web-based ChatGPT to process textual data from a multitude of research articles, OpenAI's *GPT-3.5-turbo*, which is identical to the one underpinning the ChatGPT product, facilitates a more efficient approach, as it incorporates an application programming interface (API), enabling batch processing of text from an extensive array of articles. This is achieved through iterative context and prompt submissions to ChatGPT, followed by the collection of its responses (Supporting Information, Section S3.4).

Specifically, our approach involves having ChatGPT create Python scripts for parsing academic papers, generating prompts, executing text processing through Processes 1, 2, and 3, and collating the responses into cleaned, tabulated data (Supporting Information, Figures S28–S39). Traditionally, such a process could necessitate substantial coding experience and could be time-consuming. However, we leverage the code generation capabilities of ChatGPT to establish Processes 1, 2, and 3 for batch processing using OpenAI's APIs, namely, *gpt-3.5-turbo* and *text-embedding-ada-002*. In essence, researchers only need to express their requirements for each model in natural language, specifying inputs and desired outputs, and ChatGPT will generate the appropriate Python code (Supporting Information, Section S3.5). This code can be copied, pasted, and executed in the relevant environment. Notably, even in the event of an error, ChatGPT, especially when equipped with the GPT-4 model, can assist in code revision. We note that while coding assistance from ChatGPT may not be necessary for those with coding experience, it does provide an accessible platform for individuals lacking such experience to engage in the process. Given the simplicity and straightforwardness of the logic involved in Processes 1, 2, and 3, ChatGPT-generated Python code exhibits minimal errors and significantly accelerates the programming process.

ChatGPT also aids in entity resolution after text mining (Figure 4). This step involves standardizing data formats including units, notation, and compound representations. For each task, we designed a specific prompt for ChatGPT to handle data directly or a specialized Python code generated by ChatGPT. More details on designing prompts to handle different synthesis parameters are available in a cookbook style in the Supporting Information, Section S4. In simpler cases, ChatGPT can directly handle conversions such as time and reaction temperature. For complex calculations, we take advantage of ChatGPT in generating Python code. For instance, to calculate the molar mass of each metal source, ChatGPT can generate the appropriate Python code based on the given compound formulas. For harmonizing the notation of compound pairs or mixtures, ChatGPT can standardize different notations to a unified format, facilitating subsequent data processing.

To standardize compound representations, we employed the simplified molecular input line-entry system (SMILES). We faced challenges with some synthesis procedures, where only abbreviations were provided. To overcome this, we designed prompts for ChatGPT to search for the full names of the given abbreviations. We then created a dictionary linking each unique PubChem Compound identification number (CID) or Chemical

Abstracts Service (CAS) number to multiple full names and abbreviations and generated the corresponding SMILES code. We note that for complicated linkers or those with missing full names, inappropriate nomenclature, or nonexistent CID or CAS numbers,[26–33] manual intervention was occasionally necessary to generate SMILES codes for such chemicals (Supporting Information, Figures S50–S54). However, most straightforward cases were handled efficiently by ChatGPT's generated Python code. As a result, we achieved uniformly formatted data, ready for subsequent evaluation and utilization.

RESULTS AND DISCUSSION

Evaluation of Text Mining Performance.

We began our performance analysis by first evaluating the execution time consumption for each process (Figure 5a). As previously outlined, the ChatGPT assistant in Process 1 exclusively accepts preselected experimental sections for summarization. Consequently, Process 1 requires human intervention for the identification and extraction of the synthesis section from a paper to operate autonomously. As illustrated in Figure 5a, this process can vary in duration based on the length and structure of the document and its Supporting Information file. In our study, the complete selection procedure spanned 12 h for 228 papers, averaging around 2.5 min per paper. This period must be considered to be the requisite time for Process 1's execution. For summarization tasks, ChatGPT Chemistry Assistant demonstrated an impressive performance, taking an average of 13 s per paper. This is noteworthy considering that certain papers in the data set contained more than 20 MOF compounds, and human summarization in the traditional way without AI might consume a significantly larger duration. By accelerating the summarization process, we alleviate the burden of repetitive work and free up valuable time for researchers.

In contrast, Process 2 operates in a fully automated manner, integrating the classification and result-passing processes to the next assistant for summarization. There is no doubt that it outperforms the manual identification and summarization combination of Process 1 in terms of speed due to ChatGPT's superior text processing capabilities. Finally, Process 3, as anticipated, is the fastest due to the incorporation of section filtering powered by embedding, reducing the classification tasks and subsequently enhancing the speed. The efficiency of Process 3 can be further optimized by storing the embeddings locally as a CSV file during the first reading of a paper, which reduces the processing time by 15–20 s (28–37% faster) in subsequent readings. This provides a convenient solution in scenarios necessitating repeated readings for comparison or the extraction of diverse information.

To evaluate the accuracy of the three processes in text mining, instead of sampling, we conducted a comprehensive analysis of the entire result data set. In particular, we manually wrote down the ground truth for all 11 parameters for approximately 800 compounds reported in all papers across the three processes, which was used to judge the text mining output. This involved the grading of nearly 26 000 synthesis parameters by us. Each synthesis parameter was assigned one of three labels: true positive (TP, correct identification of synthesis parameters by ChatGPT), false positive (FP, incorrect assignment of a compound to the wrong synthesis parameter or extraction of irrelevant information), or false negative (FN, failure of ChatGPT to extract some synthesis parameters). Notably, a

special rule for assigning labels on modulators, most of which were anticipated to be acid and base, was introduced to accommodate the neutral solvents in a mixed-solvent system due to the inherent challenges in distinguishing between cosolvents and modulators. For instance, in a DMF:H_2O = 10:1 solution, the role of H_2O becomes ambiguous. In such situations, we labeled the result as a TP if H_2O was considered to be either a solvent or modulator. However, we labeled it as FP or FN if it appeared or was absent in both solvent and modulator columns. Nevertheless, acids and bases were still classified as modulators, and if labeled as solvents, they were graded as FP.

The distribution of TP labels counted for each of the 11 synthesis parameters across all papers is presented in Figure 5b. It should be noted that not all MOF synthesis conditions necessitate the reporting of all 11 parameters; for instance, some syntheses do not involve modulators, and in such cases, we asked ChatGPT to assign an N/A to the corresponding column and its amount. Subsequently, we computed the precision, recall, and F1 scores for each parameter across all three processes, as illustrated in Figure 5c and d. All processes demonstrated commendable performance in identifying compound names, metal source names, linker names, modulator names, and solvent names. However, they encountered difficulties in accurately determining the quantities or volumes of the chemicals involved. Meanwhile, parameters such as the reaction temperature and reaction time, which usually have fixed patterns (e.g., units such as °C and hours, respectively), were accurately identified by all processes, resulting in high recall, precision, and F1 scores. The lowest scores were associated with the recall of solvent volumes. This is because ChatGPT often captured only one volume in mixed solvent systems instead of multiple volumes. Moreover, in some of the literature, the stock solution was used to dissolve metals and linkers, and in principle these volumes should be added to the total volume. Unfortunately, ChatGPT lacked the ability to report the volume for each portion in these cases.

Nevertheless, it should be noted that our instructions did not intend for ChatGPT to perform arithmetic operations in these cases, as the mathematical reasoning of the large language models is limited, and the diminishment of the recall scores is unavoidable. In other instances, only one exemplary synthesis condition for MOF was reported, and then for similar MOFs, the paper would state only “following similar procedures”. In such cases, while occasionally ChatGPT could duplicate conditions, most of the time it recognized solvents, the reaction temperature, and the reaction time as N/A, which was graded as a FN, thus reducing the recall scores across all processes.

Despite these irregularities, which were primarily attributable to informal synthesis reporting styles, the precision, recall, and F1 scores for all three processes remained impressively high, with less than 9.8% of NP and 0 cases of hallucination detected by human evaluators. We further calculated the average and standard deviation of each process on precision, recall, and F1 scores, as shown in Figure 5c. By considering and averaging precision, recall, and F1 scores across the 11 parameters, given their equal importance in evaluating the overall performance of the process, we found that all three processes achieved impressive precision (>95%), recall (>90%), and F1 scores (>92%).

The performance metrics of Process 1 substantiated our hypothesis that ChatGPT excels in summarization tasks. Upon comparing the performance of Processes 2 and 3—both of which are fully automated paper-reading systems capable of generating data sets from PDFs with a single click—we observed that Process 2, by meticulously examining every paragraph across all papers, ensures high precision and recall by circumventing the omission of any synthesis paragraphs or the extraction of incorrect data from irrelevant sections. Conversely, while Process 3's accuracy is marginally lower than that of Process 2, it provides a significant reduction in processing time, thus enabling faster paper reading while maintaining acceptable accuracy, courtesy of its useful filtration process.

To the best of our knowledge, these scores surpass most of those of the other models in text mining in the MOF-related domain.[11,13,14,34,35] Notably, the entire workflow, established via code and programs generated from ChatGPT, can be assembled by one or two researchers with only basic coding proficiency in a period of as brief as 1 week while maintaining remarkable performance. The successful establishment of this innovative ChatGPT Chemistry Assistant workflow, including the ChemPrompt Engineering system, which harnesses AI for processing chemistry-related tasks, promises to significantly streamline scientific research. It liberates researchers from routine laborious work, enabling them to concentrate on more focused and innovative tasks. Consequently, we anticipate that this approach will catalyze potentially revolutionary shifts in research practices through the integration of AI-powered tools.

Prediction Modeling of MOF Synthesis Outcomes.

Given the large quantity of synthesis conditions obtained through our ChatGPT-based text mining programs, our aim is to utilize these data to investigate, comprehend, and predict the crystallization conditions of a material of interest. Specifically, our goal was to determine the crystalline state based on synthesis conditions; we seek to discern which synthesis conditions will yield MOFs in the form of single crystals and which conditions are likely to yield nonsingle crystal forms of MOFs, such as microcrystalline powder or solids.

With this objective in mind, we identified the need for a label signifying the crystalline state of the resulting MOF for each synthesis condition, thereby forming a target variable for prediction. Fortunately, nearly all research papers in the MOF field consistently include the description of crystal morphological characteristics such as the color and shape of as-synthesized MOFs (e.g., yellow needle crystals, red solid, sky-blue powdered product). This facilitated rerunning our processes with the same synthesis paragraphs as input and modifying the prompt to instruct ChatGPT to extract the description of reaction products, summarizing and categorizing them (Supporting Information, Figures S23 and S47). The final label for each condition will be either single-crystal (SC) or polycrystalline (P), and our objective is to construct a machine learning model capable of accurately predicting whether a given condition will yield SC or P. Furthermore, we recognized that the crystallization process is intrinsically linked to the synthesis method (e.g., vapor diffusion, solvothermal, conventional, or microwave-assisted). Thus, we incorporated an additional synthesis variable, the "synthesis method", to categorize each synthesis condition into four distinct groups. Extracting the reaction type variable for each synthesis condition can be

achieved using the same input but a different few-shot prompt to guide our ChatGPT-based assistants for classification and summarization, subsequently merging this data with the existing data set. This process parallels the method for obtaining the MOF crystalline state outcomes, and both processes can be unified in a single prompt. Moreover, as the name of the MOF is a user-defined term and does not influence the synthesis result, we excluded this variable for the purposes of prediction modeling.

After unifying and organizing the data to incorporate 11 synthesis parameter variables and 1 synthesis outcome target variable, we designed respective descriptors for each synthesis parameter capable of robustly representing the diversity and complexity of the synthesis conditions and facilitating the transformation of these variables into features suitable for machine learning algorithms. A total of six sets of chemical descriptors were formulated for the metal node(s), linker(s), modulator(s), solvent(s), their respective molar ratios, and the reaction condition(s), aligning with the extracted synthesis parameters (Supporting Information, Section S5).[36–40] These MOF-tailored hierarchical descriptors have been previously shown to perform well in various prediction tasks.[13,41] To distill the most pertinent features and streamline the model, a recursive feature elimination (REF) with 5-fold cross-validation was performed on 80% of the total data. The rest was preserved as a held-out set unseen during the learning process for independent evaluation (Figure 6a). This down-selection process reduced the number of descriptors from 70 to 33, thereby preserving comparative model performance on the held-out set while removing the noninformative features that can lead to overfitting (Supporting Information, Section S5).

Subsequently, we constructed a machine learning model to train for synthesis conditions to predict whether a given synthesis condition can yield single crystals. A binary classifier was trained based on a random forest model (Supporting Information, Section S5). The random forest (RF) is an ensemble of decision trees whose independent predictions are max voted in the classification case to arrive at the more precise prediction.[42] In our study, we trained an RF classifier to predict crystalline states from synthesis parameters, given its ability to work with both continuous and categorical data, its advantage in ranking important features toward prediction, its robustness against noisy data,[43] and its demonstrated efficacy in various chemistry applications such as chemical property estimation,[44–47] spectroscopic analysis,[48–51] and material characterization and discovery.[52]

The dimension-reduced data was randomly divided into different training sizes; for each train test split, optimal hyperparameters, in particular, the number of tree estimators and minimum samples required for leaf split, were determined with 5-fold cross validation of the training set. Model performance was gauged in terms of class weighted accuracy, precision, recall, and F1 score over 10 runs on the held-out set and test set (Figure 6b and Supporting Information, Figure S64). The model converged to an average accuracy of 87% and an F1 score of 92% on the held-out set, indicating a reasonable performance in the presence of the imbalanced classification challenge.

Following the creation of the predictive model, our objective was to apply this model for descriptor analysis to illuminate the factors impacting MOF crystalline outcomes. This aids in discerning which features in the synthesis protocol are more crucial in determining

whether a synthesis condition will yield MOF single crystals. Although the random forest model is not inherently interpretable, we probed the relative importance of the descriptors used in building the model. One potential measure of a descriptor's importance is the percent decrease in the model's accuracy score when values for that descriptor are randomly shuffled and the model is retrained. We found that among the descriptors involved, the top 10 most influential descriptors are key in predicting MOF crystallization outcomes (Figure 6c). In fact, these descriptors broadly align with the chemical intuition and our understanding on MOF crystal growth.[53,54] For example, the descriptors related to the stoichiometry of the MOF synthesis, namely, the modulator to metal ratio, solvent to metal ratio, and linker to metal ratio, take precedence in the ranking. These descriptors reflect the vital role of precise stoichiometric control in MOF crystal formation and directly impact the crystallization process, playing critical roles in determining the quality and morphology of the MOF crystals.

Following closely is the descriptor "time", and it highlights the significant role of reaction duration in the crystallization process. Additionally, the "metal valence" descriptor emphasizes the key role of the nature and reactivity of the metal ions used in MOF synthesis. The valence directly influences the secondary building units (SBUs) and the final crystalline state of the MOF. In the meantime, descriptors related to the molecule and the linker can impact the kinetics of the synthesis, influencing the orderliness of crystal growth. Together, this result provides a greater understanding of the crucial factors affecting the crystallization of MOFs and will aid in the design and optimization of synthesis conditions for the targeted preparation of single-crystal or polycrystalline MOFs (Figure 6d).

Interrogating the Synthesis Data Set via a Chatbot.

Having utilized text mining techniques to construct a comprehensive MOF Synthesis Data set, our aim was to leverage this resource to its fullest potential. To enhance data accessibility and aid in the interpretation of its intricate contents, we embarked on a journey to convert this data set into an interactive and user-friendly dialogue system, which effectively converts the data set to dialogue. The resulting chatbot is part of the umbrella concept of the ChatGPT Chemistry Assistant thus serving as a reliable and fact-based assistant in chemistry, proficient in addressing a broad spectrum of queries pertaining to chemical reactions, in particular, MOF synthesis. Unlike typical and more general web-based ChatGPT provided by OpenAI, it may suffer from limitations such as the inability to access the most recent data and a propensity for hallucinatory errors. This chatbot is grounded firmly in the factual data contained within the MOF synthesis data set from text mining and is engineered to ensure that responses during conversations are based on accurate information and synthesis conditions derived from text mining the literature (Supporting Information, Section S6).

In particular, to construct the chemistry chatbot, our initial step was the creation of distinct entries corresponding to each MOF we identified from the text mining, which encompasses a comprehensive array of synthesis parameters, such as the reaction time, temperature, metal, and linker, among others, using the data set we have. Recognizing the value of bibliographic context, we compiled a list of paper information, such as

authors, DOI, and publication years, collated from the Web of Science, into each section (Supporting Information, Table S3). Subsequently, we generated embeddings for each of these information cards of different compounds, thereby constructing an embedding data set (Figure 7). When a user asks a question, if it is the first query, the system first navigates to the embedding data set to locate the most relevant information card using the question's embedding, which is based on a similarity score calculation and is similar to the foundation of Process 3 in text mining. The information on the highest-ranking entry is then dispatched to the prompt engineering module of the MOF chatbot, guiding it to construct responses centered solely around the given synthesis information.

To mitigate the possibility of hallucination, the chatbot is programmed to refrain from addressing queries that fall outside the scope of the data set. Instead, it encourages the user to rephrase the question (Supporting Information, Figure S69). It is worth noting that, following the initial query, the chatbot "memorizes" the conversation context by being presented with the context of prior interactions between the user and itself. This includes the synthesis context and paper information identified from the initial query, ensuring that the answers to subsequent queries are also based on factual information from the data set. Consequently, this strategy guarantees that responses to ensuing queries are contextually accurate, being grounded in the facts outlined in the synthesis data set and corresponding paper information (Figure 7 and Supporting Information, Figures S71–S74).

By virtue of its design, the chatbot addresses the challenge of enhancing data accessibility and interpretation. It accomplishes this by delivering synthesis parameters and procedures in a clear and comprehensible manner. Furthermore, it ensures data integrity and traceability by providing DOI links to the original papers, guiding users directly to the source of information. This functionality is particularly beneficial for newcomers to the field. By leveraging ChatGPT's general knowledge base, they can receive guided instructions through the synthesis process, even when faced with a procedure in a journal that is ambiguously or vaguely described. In this case, the user can consult ChatGPT to "chat with the paper" for a more precise explanation, thereby simplifying the learning process and facilitating a more efficient understanding of complex synthesis procedures. This capability fosters independent learning and expedites the comprehension of intricate synthesis procedures, reinforcing ChatGPT's role as a valuable assistant in the field of chemistry research.

Exploring Adaptability and Versatility in Large Language Models.

The adaptability of LLM-based programs, a hallmark feature distinguishing them from traditional NLP programs, lies in their inherent ability to modify search targets or tasks simply by adjusting the input prompt. Whereas traditional NLP models may necessitate a complete overhaul of rules and coding in the event of task modifications, programs powered by ChatGPT and some other LLMs utilize a more intuitive approach. A simple change in narrative language within the prompt can adequately steer the model toward the intended task, obviating the need for elaborate code adjustments.

However, we recognize limitations within the current workflow, particularly concerning token limitations. Research articles for text mining were parsed into short snippets due to the 4096 token limit from *GPT-3.5-turbo*, since longer research articles can extend to 20 000–40

000 tokens. This fragmentation may inadvertently result in the undesirable segmentation of the synthesis paragraphs or other sections containing pertinent information. To alleviate this, we envision that a large language model that can process higher token memory[61,62] such as *GPT-4–32K* (OpenAI) or *Claude-v1* (Anthropic) will be very helpful since each time, it reads the entire paper rather than just sections, which can further increase its accuracy by avoiding undesirable segmentation of the synthesis paragraph or other targeted paragraphs containing information. Longer reading capabilities will also have the added benefit of reducing the number of tokens used in repeated questions, thus enhancing processing times. As we continue to refine our workflow, we believe that there are additional opportunities for improvement. For instance, parts of the fixed prompt could be more concise to save tokens, and the examples in the few-shot prompt can be further optimized to reduce the total tokens. Given that each paper may have around 100 segments, such refinements could dramatically reduce time and costs, particularly for classification and summarization tasks, which must process every section with the same fixed prompt, especially for few-shot instructions.

Furthermore, language versatility, a crucial aspect in the realm of text mining, is seamlessly addressed by LLMs. Traditional NLP models, trained in a specific language, often struggle when the task requires processing text data in another language. For example, if the model is trained on English data, it may require substantial adjustments or even a complete rewrite to process text data in Arabic, Chinese, French, German, French, Japanese, Korean, and some other languages. However, with LLMs that can handle multiple languages, such as ChatGPT, we showed that researchers just need to slightly alter the instructions or prompts to achieve the goal, without the necessity of substantial code modifications (Supporting Information, Figures S55–S58).

The adaptable nature of LLMs can further extend their versatility in handling diverse tasks. We demonstrated how prompts can be changed to direct ChatGPT to parse and summarize different types of information from the same pool of research articles. For instance, with minor modification of the prompts, we show that our ChatGPT Chemistry Assistants have the potential to be instructed to summarize diverse information such as thermal stability, BET surface area, CO_2 uptake, crystal parameters, water stability, and even MOF structure or topology (Supporting Information, Section S4). This adaptability was previously a labor-intensive process, requiring experienced specialists to manually collect or establish training sets for text mining each type of information.[11,13,35,41,63–66]

Moreover, the utility of this approach can benefit the broader chemistry domain: it is capable of not only facilitating data mining in research papers addressing MOF synthesis but also extending it to all chemistry papers with the accorded modifications. By fine-tuning the prompt, the ChatGPT Chemistry Assistant can effectively extract and tabulate data from diverse fields, such as organic synthesis, biochemistry preparations, perovskite preparations, polymer synthesis, and more. This capability underscores the versatility of the ChatGPT-based assistant, not only in terms of subject matter but also in terms of the level of detail it can handle. In the event that key parameters for data extraction are not explicitly defined, ChatGPT can be prompted to suggest parameters based on its trained understanding of the text. This level of adaptability and interactivity is unparalleled in traditional NLP models, highlighting a key advantage of the ChatGPT approach. The shift

from a code-intensive approach to a natural language instruction approach democratizes the process of data mining, making it accessible even to those with less coding expertise, making it an innovative and powerful solution for diverse data mining challenges.

CONCLUDING REMARKS

Our research has successfully demonstrated the potential of LLMs, particularly GPT models, in the domain of chemistry research. We presented a ChatGPT Chemistry Assistant that includes three different but connected approaches to text mining with ChemPrompt Engineering: Process 3 is capable of conducting search and filtration, Processes 2 and 3 classify synthesis paragraphs, and Processes 1, 2, and 3 are capable of summarizing synthesis conditions into structured data sets. Enhanced by three fundamental principles of prompt engineering specific to chemistry text processing, coupled with the interactive prompt refinement strategy, the ChatGPT-based assistant has substantially advanced the extraction and analysis of the MOF synthesis literature, with precision, recall, and F1 scores exceeding 90%.

We elucidated two crucial insights from the data set of synthesis conditions. First, the data can be employed to construct predictive models for reaction outcomes, which shed light on the key experimental factors that influence the MOF crystallization process. Second, it is possible to create an MOF chatbot that can provide accurate answers based on text mining, thereby improving access to the synthesis data set and achieving a data-to-dialogue transition. This investigation illustrates the potential for rapid advancement inherent in ChatGPT and other LLMs as a proof of concept.

On a fundamental level, this study provides guidance on interacting with LLMs to serve as AI assistants for chemists, accelerating research with minimal prerequisite coding expertise and thus bridging the gap between chemistry and the realms of computational and data science more effectively. Through interaction and chatting, the code and design of experiments can be modified, democratizing data mining and enhancing the landscape of scientific research. Our work sets a foundation for further exploration and application of LLMs across various scientific domains, paving the way for a new era of AI-assisted chemistry research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

Z.Z. extends special gratitude to Jiayi Weng (OpenAI) for valuable discussions on harnessing the potential of ChatGPT. In addition, Z.Z. acknowledges the inspiring guidance and input from Kefan Dong (Stanford University), Long Lian (University of California, Berkeley), and Yifan Deng (Carnegie Mellon University), all of whom contributed to shaping the study's design and enhancing the performance of ChatGPT. We express our appreciation to Dr. Nakul Rampal from the Yaghi laboratory for insightful discussions. Our gratitude is also extended for the financial support received from the Defense Advanced Research Projects Agency (DARPA) under contract HR0011–21-C-0020. O.Z. acknowledges funding and extends thanks for the support provided by the National Institutes of Health (NIH) under grant 5R01GM127627–04. Additionally, Z.Z. is grateful for the financial support received through a Kavli ENSI Graduate Student Fellowship and the Bakar Institute of Digital Materials for the

Planet (BIDMaP). This work is independently developed by the University of California, Berkeley research team and not affiliated, endorsed, or sponsored by OpenAI.

REFERENCES

(1). Yaghi OM; O'Keeffe M; Ockwig NW; Chae HK; Eddaoudi M; Kim J Reticular synthesis and the design of new materials. Nature 2003, 423 (6941), 705–714. [PubMed: 12802325]

(2). Matlin SA; Mehta G; Hopf H; Krief A The role of chemistry in inventing a sustainable future. Nat. Chem. 2015, 7 (12), 941–943. [PubMed: 26587703]

(3). Bubeck S; Chandrasekaran V; Eldan R; Gehrke J; Horvitz E; Kamar E; Lee P; Lee YT; Li Y; Lundberg S Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv DOI: 10.48550/arXiv.2303.12712 (accessed 2023-04-13).

(4). Aspuru-Guzik A; Lindh R; Reiher M The matter simulation (r) evolution. ACS Cent. Sci. 2018, 4 (2), 144–152. [PubMed: 29532014]

(5). Chen H; Engkvist O; Wang Y; Olivecrona M; Blaschke T The rise of deep learning in drug discovery. Drug Discovery Today 2018, 23 (6), 1241–1250. [PubMed: 29366762]

(6). Kaspar C; Ravoo B; van der Wiel WG; Wegner S; Pernice W The rise of intelligent matter. Nature 2021, 594 (7863), 345–355. [PubMed: 34135518]

(7). Gómez-Bombarelli R; Wei JN; Duvenaud D; Hernández-Lobato JM; Sánchez-Lengeling B; Sheberla D; Aguilera-Iparraguirre J; Hirzel TD; Adams RP; Aspuru-Guzik A Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. 2018, 4 (2), 268–276. [PubMed: 29532027]

(8). Firat M What ChatGPT means for universities: Perceptions of scholars and students. J. Appl. Learn. Teach. 2023, 6(1), DOI: 10.37074/jalt.2023.6.1.22.

(9). Lyu H; Ji Z; Wuttke S; Yaghi OM Digital reticular chemistry. Chem. 2020, 6 (9), 2219–2241.

(10). Jensen Z; Kim E; Kwon S; Gani TZ; Román-Leshkov Y; Moliner M; Corma A; Olivetti E A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. ACS Cent. Sci. 2019, 5 (5), 892–899. [PubMed: 31139725]

(11). Park S; Kim B; Choi S; Boyd PG; Smit B; Kim J Text mining metal-organic framework papers. J. Chem. Inf. Model. 2018, 58 (2), 244–251. [PubMed: 29227671]

(12). He T; Sun W; Huo H; Kononova O; Rong Z; Tshitoyan V; Botari T; Ceder G Similarity of precursors in solid-state synthesis as text-mined from scientific literature. Chem. Mater. 2020, 32 (18), 7861–7873.

(13). Park H; Kang Y; Choe W; Kim J Mining Insights on Metal-Organic Framework Synthesis from Scientific Literature Texts. J. Chem. Inf. Model. 2022, 62 (5), 1190–1198. [PubMed: 35195419]

(14). Luo Y; Bag S; Zaremba O; Cierpka A; Andreo J; Wuttke S; Friederich P; Tsotsalas M MOF synthesis prediction enabled by automatic data mining and machine learning. Angew. Chem., Int. Ed. 2022, 61 (19), No. e202200242.

(15). Brown T; Mann B; Ryder N; Subbiah M; Kaplan JD; Dhariwal P; Neelakantan A; Shyam P; Sastry G; Askell A Language models are few-shot learners. NIPS 2020, 33, 1877–1901.

(16). Radford A; Wu J; Child R; Luan D; Amodei D; Sutskever I Language models are unsupervised multitask learners. OpenAI blog 2019, 1 (8), 9.

(17). Radford A; Narasimhan K; Salimans T; Sutskever I, Improving language understanding by generative pre-training. Preprint2018.

(18). Jablonka KM; Schwaller P; Ortega-Guerrero A; Smit B Is GPT-3 all you need for low-data discovery in chemistry? ChemRxiv DOI: 10.26434/chemrxiv-2023-fw8n4 (accessed 2023-02-14).

(19). Moghadam PZ; Li A; Wiggin SB; Tao A; Maloney AG; Wood PA; Ward SC; Fairen-Jimenez D Development of a Cambridge Structural Database subset: a collection of metal-organic frameworks for past, present, and future. Chem. Mater. 2017, 29 (7), 2618–2625.

(20). Chung YG; Camp J; Haranczyk M; Sikora BJ; Bury W; Krungleviciute V; Yildirim T; Farha OK; Sholl DS; Snurr RQ Computation-ready, experimental metal-organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. Chem. Mater. 2014, 26 (21), 6185–6192.

(21). Chung YG; Haldoupis E; Bucior BJ; Haranczyk M; Lee S; Zhang H; Vogiatzis KD; Milisavljevic M; Ling S; Camp JS Advances, updates, and analytics for the computation-ready, experimental metal-organic framework database: CoRE MOF 2019. J. Chem. Eng. Data 2019, 64 (12), 5985–5998.

(22). Mikolov T; Chen K; Corrado G; Dean J Efficient estimation of word representations in vector space. arXiv DOI: 10.48550/arXiv.1301.3781 (accessed 2013-09-07).

(23). Le Q; Mikolov T In Distributed Representations of Sentences and Documents; International Conference on Machine Learning, PMLR: 2014; pp 1188–1196.

(24). Mikolov T; Sutskever I; Chen K; Corrado GS; Dean J Distributed representations of words and phrases and their compositionality. NIPS 2013, 26.

(25). Kusner M; Sun Y; Kolkin N; Weinberger K From Word Embeddings to Document Distances; International Conference on Machine Learning, PMLR: 2015; pp 957–966.

(26). Gong W; Xie H; Idrees KB; Son FA; Chen Z; Sha F; Liu Y; Cui Y; Farha OK Water sorption evolution enabled by reticular construction of zirconium metal-organic frameworks based on a unique [2.2] paracyclophane scaffold. J. Am. Chem. Soc. 2022, 144 (4), 1826–1834. [PubMed: 35061394]

(27). Hanikel N; Kurandina D; Chheda S; Zheng Z; Rong Z; Neumann SE; Sauer J; Siepmann JI; Gagliardi L; Yaghi OM MOF Linker Extension Strategy for Enhanced Atmospheric Water Harvesting. ACS Cent. Sci. 2023, 9 (3), 551–557. [PubMed: 36968524]

(28). Liu T-F; Feng D; Chen Y-P; Zou L; Bosch M; Yuan S; Wei Z; Fordham S; Wang K; Zhou H-C Topology-guided design and syntheses of highly stable mesoporous porphyrinic zirconium metal-organic frameworks with high surface area. J. Am. Chem. Soc. 2015, 137 (1), 413–419. [PubMed: 25495734]

(29). Bloch ED; Murray LJ; Queen WL; Chavan S; Maximoff SN; Bigi JP; Krishna R; Peterson VK; Grandjean F; Long GJ Selective binding of O2 over N2 in a redox-active metal-organic framework with open iron (II) coordination sites. J. Am. Chem. Soc. 2011, 133 (37), 14814–14822. [PubMed: 21830751]

(30). Furukawa H; Go YB; Ko N; Park YK; Uribe-Romo FJ; Kim J; O'Keeffe M; Yaghi OM Isoreticular expansion of metal-organic frameworks with triangular and square building units and the lowest calculated density for porous crystals. Inorg. Chem. 2011, 50 (18), 9147–9152. [PubMed: 21842896]

(31). Zheng Z; Rong Z; Iu-Fan Chen O; Yaghi OM Metal-Organic Frameworks with Rod Yttrium Secondary Building Units. Isr. J. Chem. 2023, No. e202300017.

(32). Reinsch H; van der Veen MA; Gil B; Marszalek B; Verbiest T; De Vos D; Stock N Structures, sorption characteristics, and nonlinear optical properties of a new series of highly stable aluminum MOFs. Chem. Mater. 2013, 25 (1), 17–26.

(33). Hu Z; Pramanik S; Tan K; Zheng C; Liu W; Zhang X; Chabal YJ; Li J Selective, sensitive, and reversible detection of vapor-phase high explosives via two-dimensional mapping: A new strategy for MOF-based sensors. Cryst. Growth Des. 2013, 13 (10), 4204–4207.

(34). Glasby LT; Gubsch K; Bence R; Oktavian R; Isoko K; Moosavi SM; Cordiner JL; Cole JC; Moghadam PZ DigiMOF: A Database of Metal-Organic Framework Synthesis Information Generated via Text Mining. Chem. Mater. 2023, 35, 4510. [PubMed: 37332681]

(35). Nandy A; Duan C; Kulik HJ Using machine learning and data mining to leverage community knowledge for the engineering of stable metal-organic frameworks. J. Am. Chem. Soc. 2021, 143 (42), 17535–17547. [PubMed: 34643374]

(36). Shannon RD Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Crystallogr. A 1976, 32 (5), 751–767.

(37). Haynes WM CRC Handbook of Chemistry and Physics; CRC Press: Boca Raton, FL, 2016.

(38). Pauling L The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. J. Am. Chem. Soc 1932, 54 (9), 3570–3582.

(39). Nguyen KT; Blum LC; Van Deursen R; Reymond JL Classification of organic molecules by molecular quantum numbers. ChemMedChem. 2009, 4 (11), 1803–1805. [PubMed: 19774591]

(40). Deursen R. v.; Blum LC; Reymond J-L A searchable map of PubChem. J. Chem. Inf. Model. 2010, 50 (11), 1924–1934. [PubMed: 20945869]

(41). Batra R; Chen C; Evans TG; Walton KS; Ramprasad R Prediction of water stability of metal-organic frameworks using machine learning. Nat. Mach 2020, 2 (11), 704–710.

(42). Ho TK Random Decision Forests; Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE: 1995; pp 278–282.

(43). Kaiser TM; Burger PB Error tolerance of machine learning algorithms across contemporary biological targets. Molecules 2019, 24 (11), 2115. [PubMed: 31167452]

(44). Meyer JG; Liu S; Miller IJ; Coon JJ; Gitter A Learning drug functions from chemical structures with convolutional neural networks and random forests. J. Chem. Inf. Model. 2019, 59 (10), 4438–4449. [PubMed: 31518132]

(45). Rajappan R; Shingade PD; Natarajan R; Jayaraman VK Quantitative Structure- Property Relationship (QSPR) Prediction of Liquid Viscosities of Pure Organic Compounds Employing Random Forest Regression. Ind. Eng. Chem. Res. 2009, 48 (21), 9708–9712.

(46). Kapsiani S; Howlin BJ Random forest classification for predicting lifespan-extending chemical compounds. Sci. Rep. 2021, 11, 13812. [PubMed: 34226569]

(47). Svetnik V; Liaw A; Tong C; Culberson JC; Sheridan RP; Feuston BP Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 2003, 43 (6), 1947–1958. [PubMed: 14632445]

(48). Franklin EB; Yee LD; Aumont B; Weber RJ; Grigas P; Goldstein AH Ch3MS-RF: a random forest model for chemical characterization and improved quantification of unidentified atmospheric organics detected by chromatography-mass spectrometry techniques. Atmos. Meas. Technol. 2022, 15 (12), 3779–3803.

(49). de Santana FB; Neto WB; Poppi RJ Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. Food Chem. 2019, 293, 323–332. [PubMed: 31151619]

(50). Seifert S Application of random forest based approaches to surface-enhanced Raman scattering data. Sci. Rep. 2020, 10 (1), 5436. [PubMed: 32214194]

(51). Torrisi SB; Carbone MR; Rohr BA; Montoya JH; Ha Y; Yano J; Suram SK; Hung L Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. Npj Comput. Mater. 2020, 6 (1), 109.

(52). Ahneman DT; Estrada JG; Lin S; Dreher SD; Doyle AG Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018, 360 (6385), 186–190. [PubMed: 29449509]

(53). Yaghi OM; Kalmutzki MJ; Diercks CS Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks; John Wiley & Sons: 2019.

(54). Han Y; Yang H; Guo X Synthesis Methods and Crystallization of MOFs. Synthesis Methods and Crystallization; IntechOpen: 2020; pp 1–23.

(55). Gándara F; Furukawa H; Lee S; Yaghi OM High methane storage capacity in aluminum metal-organic frameworks. J. Am. Chem. Soc. 2014, 136 (14), 5271–5274. [PubMed: 24661065]

(56). Rowsell JL; Yaghi OM Effects of functionalization, catenation, and variation of the metal oxide and organic linking units on the low-pressure hydrogen adsorption properties of metal- organic frameworks. J. Am. Chem. Soc. 2006, 128 (4), 1304–1315. [PubMed: 16433549]

(57). Li M-Y; Wang F; Zhang J Zeolitic tetrazolate-imidazolate frameworks with SOD topology for room temperature fixation of CO2 to cyclic carbonates. Cryst. Growth Des. 2020, 20 (5), 2866–2870.

(58). Zheng Z; Alawadhi AH; Yaghi OM Green Synthesis and Scale-Up of MOFs for Water Harvesting from Air. Mol. Front. J. 2023, 1–20.

(59). Köppen M; Meyer V; Ångström J; Inge AK; Stock N Solvent-dependent formation of three new Bi-metal-organic frameworks using a tetracarboxylic acid. Cryst. Growth Des. 2018, 18 (7), 4060–4067.

(60). Ma K; Cheung YH; Xie H; Wang X; Evangelopoulos M; Kirlikovali KO; Su S; Wang X; Mirkin CA; Xin JH Zirconium-Based Metal-Organic Frameworks as Reusable Antibacterial Peroxide Carriers for Protective Textiles. Chem. Mater. 2023, 35 (6), 2342–2352.

(61). Bulatov A; Kuratov Y; Burtsev MS Scaling Transformer to 1M tokens and beyond with RMT. arXiv DOI: 10.48550/arXiv.2304.11062 (accessed 2023-04-19).

(62). Dao T; Fu D; Ermon S; Rudra A; Ré C Flashattention: Fast and memory-efficient exact attention with io-awareness. NIPS 2022; Vol. 35, pp 16344–16359.

(63). Colón YJ; Gomez-Gualdron DA; Snurr RQ Topologically guided, automated construction of metal-organic frameworks and their evaluation for energy-related applications. Cryst. Growth Des. 2017, 17 (11), 5801–5810.

(64). Nandy A; Yue S; Oh C; Duan C; Terrones GG; Chung YG; Kulik HJ A database of ultrastable MOFs reassembled from stable fragments with machine learning models. Matter 2023, 6 (5), 1585–1603.

(65). Suyetin M The application of machine learning for predicting the methane uptake and working capacity of MOFs. Faraday Discuss. 2021, 231, 224–234. [PubMed: 34195741]

(66). Nandy A; Terrones G; Arunachalam N; Duan C; Kastner DW; Kulik HJ MOF Simplify, machine learning models with extracted stability data of three thousand metal-organic frameworks. Sci. Data 2022, 9 (1), 74. [PubMed: 35277533]

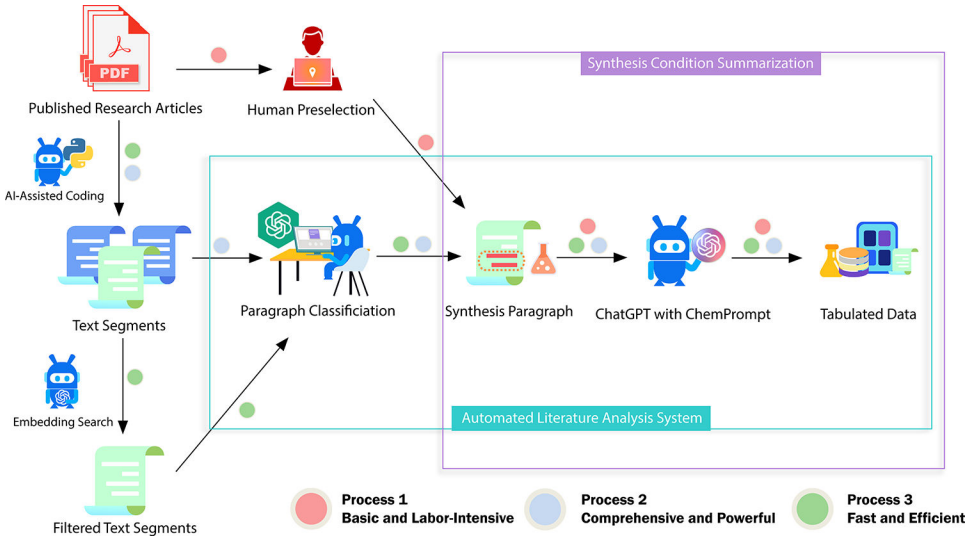


Figure 1.

Schematics of the ChatGPT Chemistry Assistant workflow having three different processes employing ChatGPT and ChemPrompt for efficient text mining and summarization of MOF synthesis conditions from a diverse set of published research articles. Each process is distinctively labeled with red, blue, and green dots. To illustrate, Process 1 is initiated with “Published Research Articles”, proceeds to “Human Preselection”, moves to the “Synthesis Paragraph”, integrates “ChatGPT with Chem-Prompt”, and culminates in “Tabulated Data”. Steps shared among multiple processes are indicated with corresponding color-coded dots. The two-snakes logo of Python is included to indicate the use of the Python programming language, with the logo’s credit attributed to the Python Software Foundation (PSF). The white or black OpenAI logo is embedded to symbolize that the process is powered by OpenAI models, with the logo’s credit acknowledged as belonging to OpenAI.

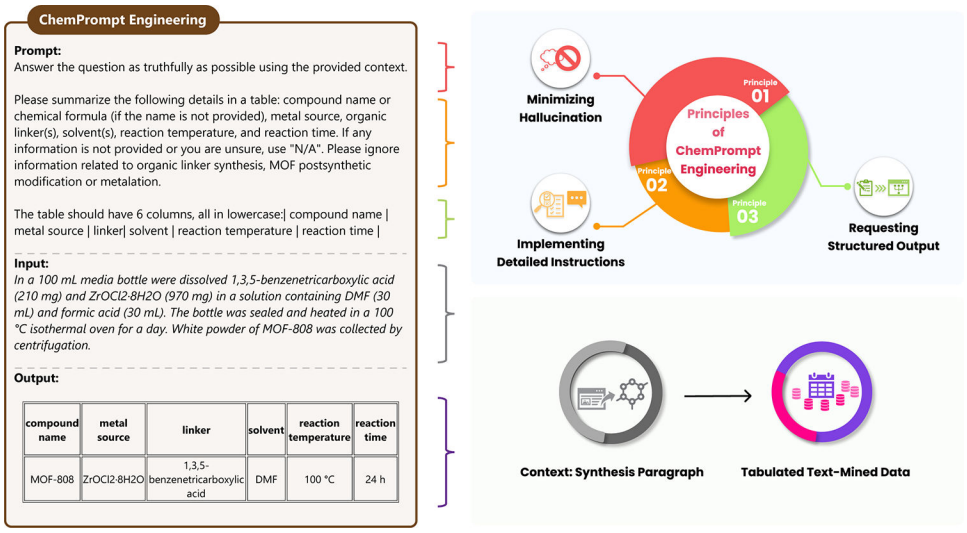


Figure 2.

Illustration of a carefully designed ChemPrompt (shown on the left) encapsulating all three fundamental principles of ChemPrompt Engineering (shown on the right). The prompt guides ChatGPT to systematically extract and summarize synthesis conditions from a specified section in a research article, organizing the data into a well-structured table.

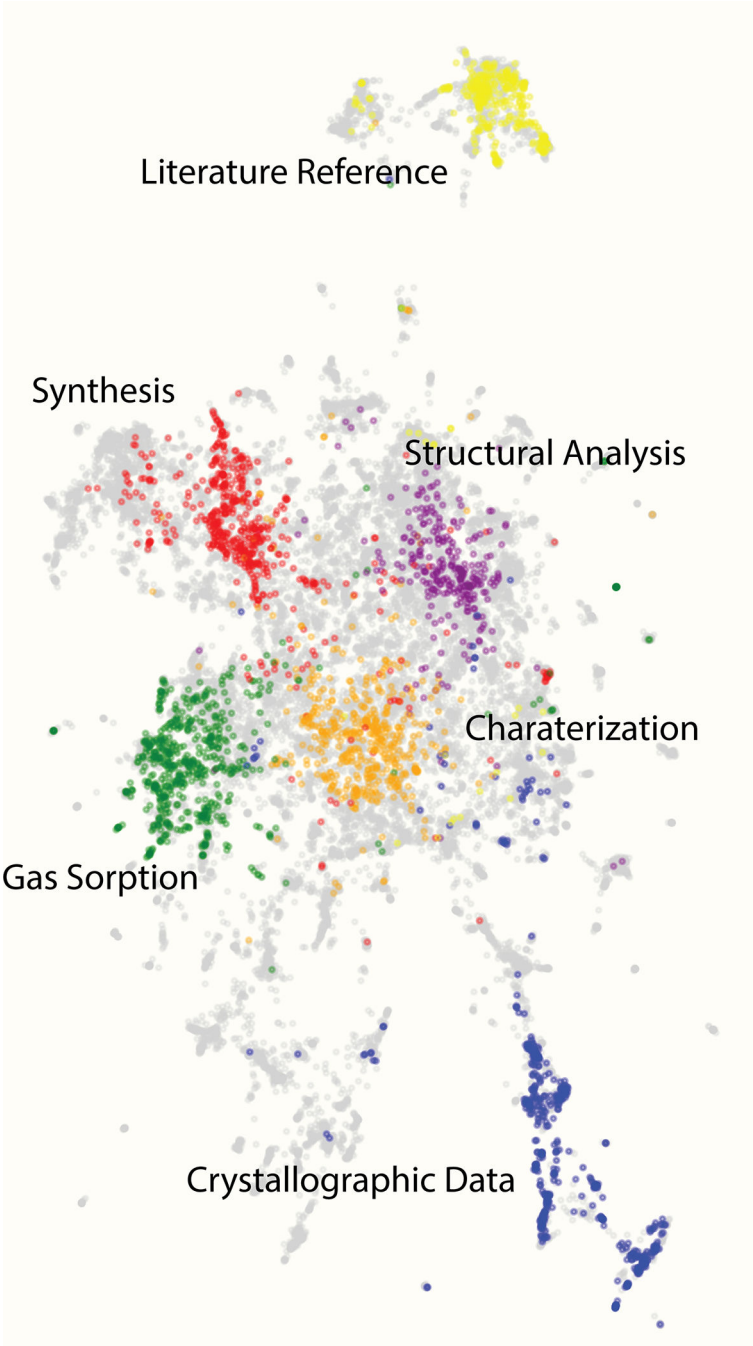


Figure 3.
Two-dimensional visualization of 18 248 text segment embeddings, with each point representing a text segment from the research articles selected. Color coding denotes thematic categories: red for synthesis, green for gas sorption, yellow for literature reference, blue for crystallographic data, purple for structural analysis, orange for characterization, and gray for other text segments not emphasized in this study.

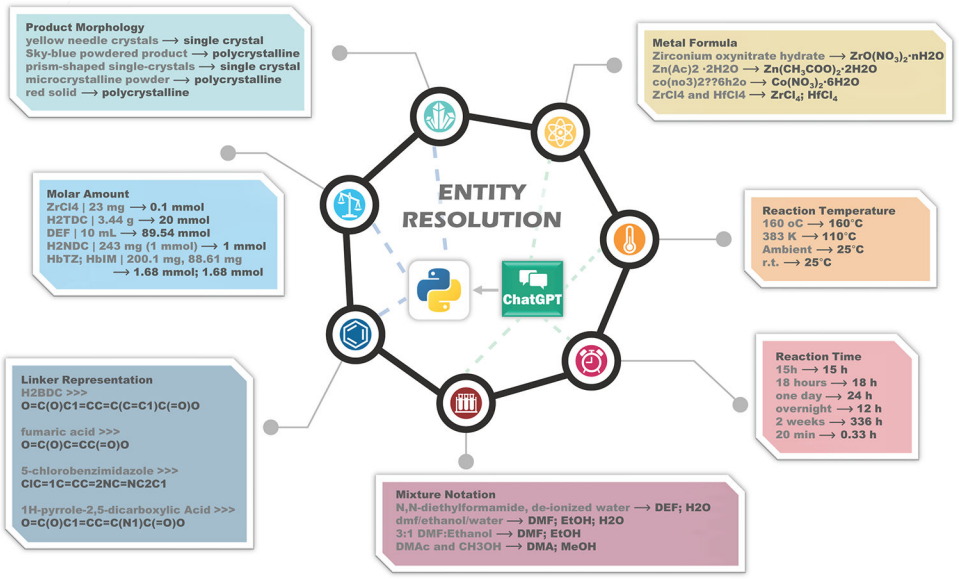


Figure 4.

Schematic representation of the diverse data unification tasks managed either directly by ChatGPT or through Python code written by ChatGPT. The figure distinguishes between simpler tasks handled directly by ChatGPT, such as standardizing chemical notation and converting time and temperature units in reactions. More complex tasks, such as matching linker abbreviations to their full names, converting these to SMILES codes, classifying product morphology, and calculating metal amounts, are accomplished via Python code generated by ChatGPT. The Python logo displayed is credited to the Python Software Foundation, while the OpenAI logo is credited to OpenAI.

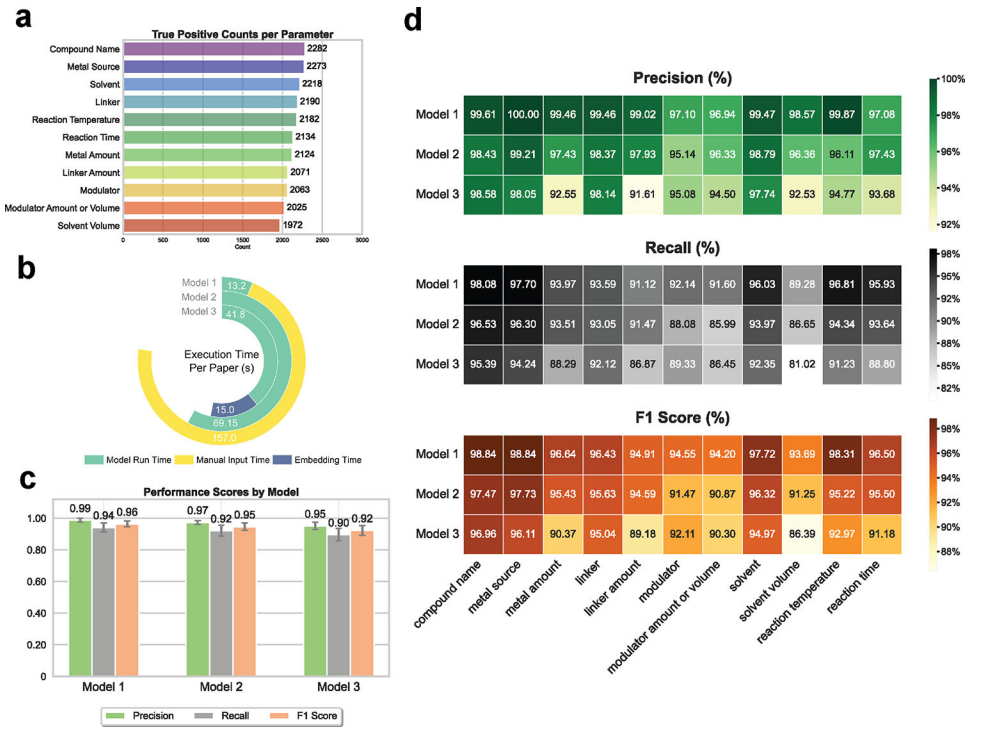


Figure 5.

Multifaceted performance analysis of ChatGPT-based text mining processes. (a) Comparison of the average execution time required by each process to read and process a single paper, highlighting their relative efficiency. (b) Distribution of true positive counts for each of the 11 synthesis parameters, derived from the cumulative results of Processes 1, 2, and 3 based on a total of 2387 synthesis conditions. Despite minor discrepancies, the counts are closely aligned, demonstrating the assistants' proficiency in effectively extracting the selected parameters. (c) Aggregate average precision, recall, and F1 scores for each process, indicating their overall accuracy and reliability. Standard deviations are represented by gray error bars in the chart. (d) Heat map illustrating the detailed percentage precision, recall, and F1 scores for each synthesis parameter across the three processes, providing a nuanced understanding of the ChatGPT-based assistants' performance in accurately identifying specific synthesis parameters.

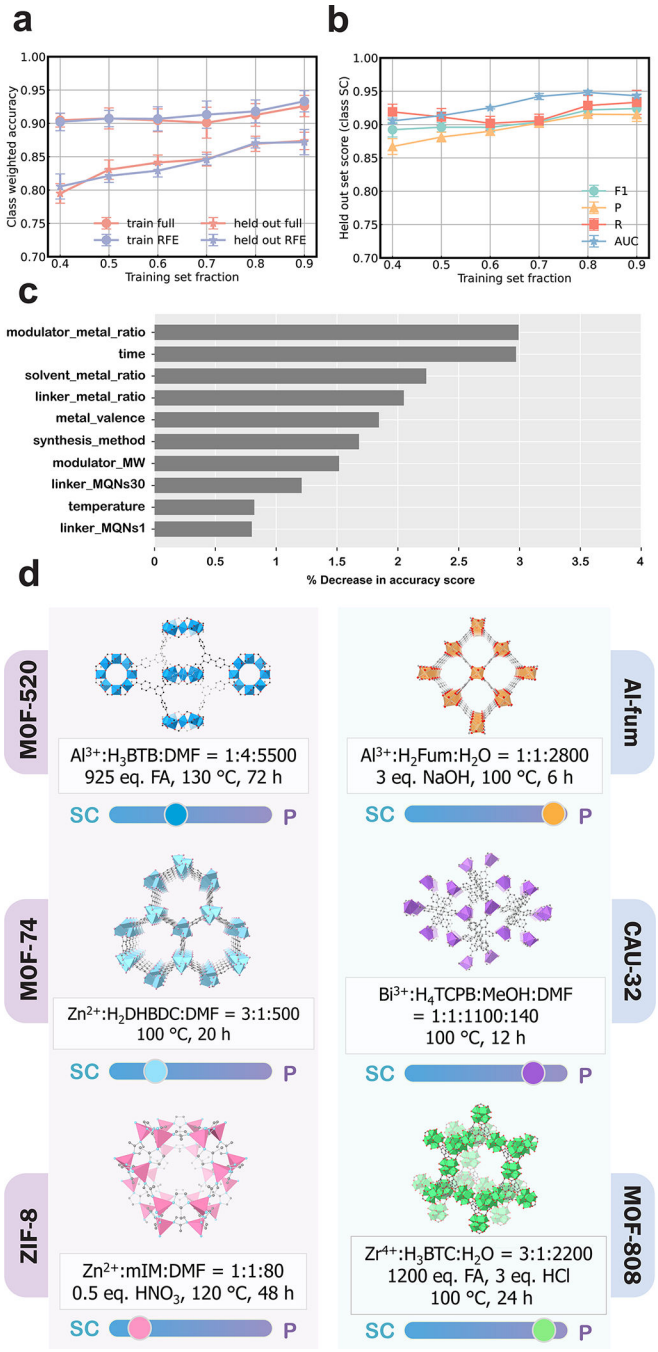


Figure 6.

Performance of the classification models in predicting the crystalline state of MOFs from synthesis. (a) Learning curves of the classifier model with 1σ standard deviation error bars. (b) Model performance evaluation through the F1 Score, Precision, Recall, and Area Under the Curve metrics. The training set fraction was in ratio to the data excluding the held-out set. (c) The 10 most significant descriptors of the trained random forest model, determined by an accuracy score increase. (d) Six examples of MOFs, MOF-520, MOF-74, ZIF-8, Al-fum, CAU-32, and MOF-808 along with their synthesis conditions derived from

the literature.[55–60] Circle positions on the bar represent the likelihood of resulting in single-crystal or polycrystalline states predicted by the model. The model's predictions for these six examples aligned with actual experimental results.

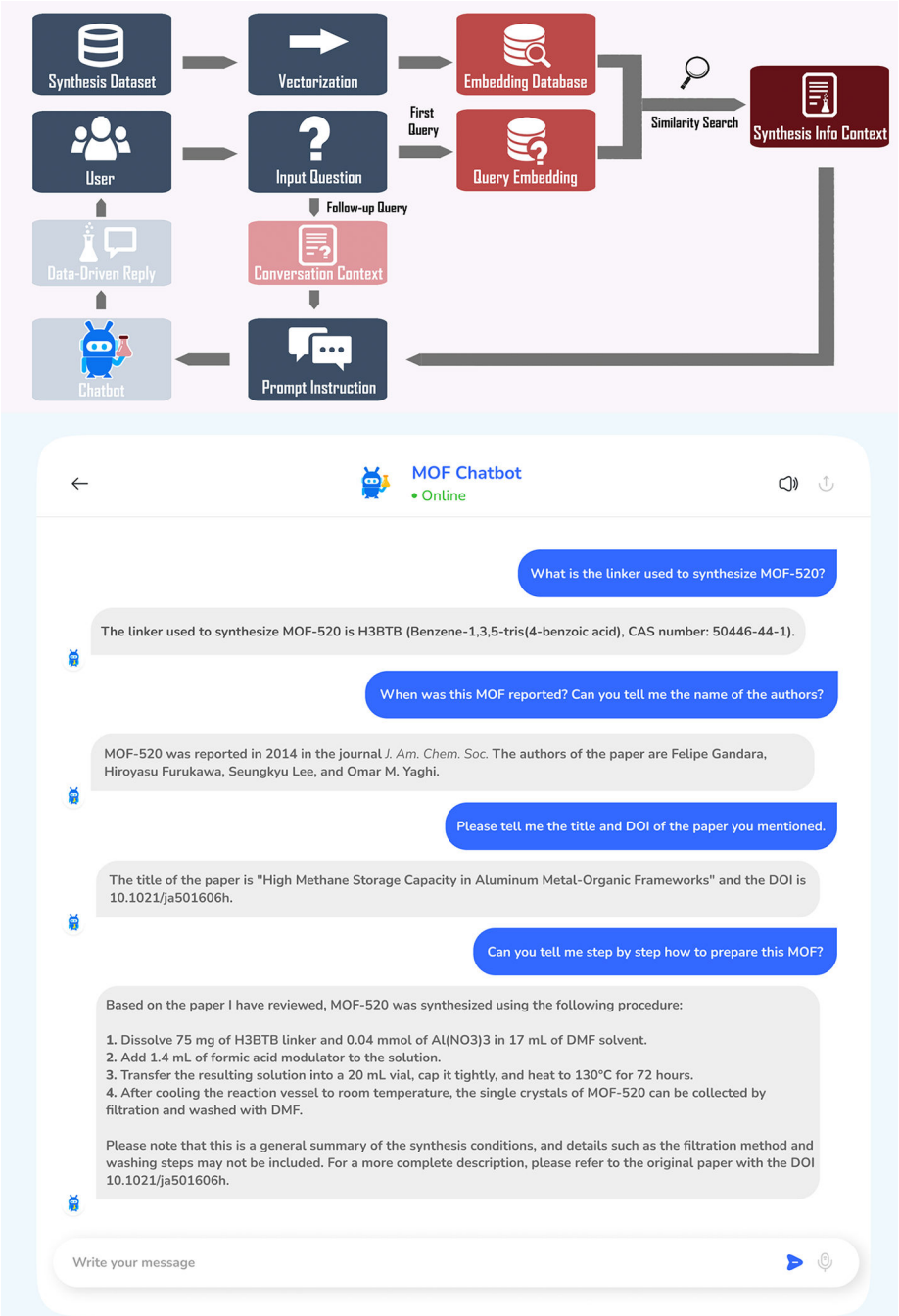


Figure 7.

Integrated workflow of the MOF chatbot transforming comprehensive synthesis data sets into contextually accurate dialogue systems and demonstration of a conversation with the data-driven chatbot. The process ensures enhanced data accessibility and interpretation and facilitates independent learning in the field of chemistry research.

Table 1.

Assessment of Hallucination in the ChatGPT Response without Prompt Engineering

Query	ChatGPT Response[a]
Which metal is used in the synthesis of MOF-5?	Zinc (Correct)
Which metal is used in the synthesis ofMOF-519?	Zirconium (Incorrect)
What is the linker used in the synthesis of MOF-99999?	I do not know (Correct)
What is the linker used in the synthesis of MOF-419?	Terephthalic acid (Incorrect)
What is the linker used in the synthesis ofZIF-8?	2-Methylimidazole (Correct)

[a]Responses are representative answers selected from a series of 100 repeated queries, followed by parenthetical indications of their correctness, which is based on the established facts concerning the respective compounds referenced in the queries.

Table 2.

Improvements in ChatGPT Response Accuracy Utilizing a Basic Prompt Engineering Strategy

Initial Query	Guided Prompt	ChatGPT Response[a]
Which metal is used in the synthesis of MOF-5?		Zinc (Correct)
Which metal is used in the synthesis ofMOF-519?		I do not know (Correct)
What is the linker used in the synthesis of MOF-99999?	If you are uncertain, please reply with “I do not know”.	I do not know (Correct)
What is the linker used in the synthesis of MOF-419?		I do not know (Correct)
What is the linker used in the synthesis of ZIF-8?		2-Methylimidazole (Correct)

[a]Responses are representative answers selected from a series of 100 repeated queries, followed by parenthetical indications of their correctness, which is based on the established facts concerning the respective compounds referenced in the queries.