Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2026, **5**, 384

# Toward smart $CO_2$ capture by the synthesis of metal organic frameworks using large language models

Hossein Mashhadimoslem, *[ab] Mohammad Ali Abdol,[a] Kourosh Zanganeh,[c] Ahmed Shafeen,[c] Encheng Liu,[d] Sohrab Zendehboudi,[e] Ali Elkamel[afg] and Aiping Yu *[ab]

This research focuses on efficiently collecting $CO_2$ adsorption data using experimental metal–organic framework (MOF) porous materials from the scientific literature, addressing the challenges related to data classification and access to MOF synthesis methods. The aim is to organize, classify, and facilitate easy access to materials science information using artificial intelligence (AI). Using advanced large language models (LLMs), we developed a systematic approach to extract and sort MOF synthesis data for $CO_2$ adsorption in a structured format. Using this method, we collected data from over 433 published experimental research papers and created a specific dataset to analyze the effects of metals, ligands, and carbon adsorption conditions on $CO_2$ uptake performance. The correlations between the material structure, such as metal types, ligands, specific surface area, pore size, pore volume, synthesis conditions, and $CO_2$ adsorption, under various process conditions were examined using the final database. We applied ChatGPT 4o mini as an AI assistant to text-mine all MOF information from different PDF file references. In addition to revealing the impact of each parameter on $CO_2$ uptake and MOF structure before synthesis, the AI analysis findings indicated which ligand and metal groups should be altered to customize the MOF structure for improved $CO_2$ capture.

Received 3rd October 2025
Accepted 18th November 2025

DOI: 10.1039/d5dd00446b

rsc.li/digitaldiscovery



## 1. Introduction

In response to growing concerns over climate change, the Intergovernmental Panel on Climate Change (IPCC) was established in 1988 to deliver regular, comprehensive scientific assessments to inform policymakers about the current understanding of climate change.[1] From 2016 to 2100, cumulative residual greenhouse gas (GHG) emissions, including $CO_2$ from fossil fuel use, are projected to range from 850 to 1150 Gt $CO_2$.[2] According to IPCC, current trends are likely to result in a global warming of 1.5 °C between 2030 and 2052, with human activity expected to contribute to an anticipated increase of 0.8–1.2 °C.[1] Growing concern over the accelerating accumulation of GHGs, primarily $CO_2$, has underscored the urgent need for effective mitigation strategies. In this context, advancements in materials science offer a promising pathway for developing novel technologies to reduce carbon emissions.[3]

Since solid adsorbent-based gas adsorption and separation methods offer the potential to be more adaptable and energy-efficient for a range of carbon capture applications, they have been extensively studied.[4] Metal oxides,[5] polymers,[6] zeolite,[7] ceria oxide,[8] porous carbon, and activated carbon[9] are the types of common adsorbents. Metal–organic frameworks (MOFs) have been extensively investigated as potent and promising $CO_2$ capture adsorbents in recent years due to their porous structure, large specific surface area ($S_{BET}$), high capacity, excellent selectivity, and structural tunability.[10] As a result, an exceptionally wide range of MOF materials can be synthesized through broad design strategies. Engineering the selection of framework components should enable precise control over the internal pore surface's affinity for $CO_2$ adsorption.[11] This will allow for the customization of MOF material properties tailored to specific $CO_2$ capture and separation processes as well as to various operating conditions. Significant progress has been

[a]*Department of Chemical Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. E-mail: hmashhadimoslem@uwaterloo.ca; aelkamel@uwaterloo.ca; aipingyu@uwaterloo.ca*

[b]*Waterloo Institute for Nanotechnology Department of Chemical Engineering, University of Waterloo, 200 University Ave. W., Waterloo, ON, N2L 3G1, Canada*

[c]*Natural Resources Canada (NRCan), CanmetENERGY-Ottawa (CE-O), 1 Haanel Dr, Ottawa, ON K1A 1M1, Canada. E-mail: kourosh.zanganeh@nrcan-rncan.gc.ca; ahmed.shafeen@nrcan-rncan.gc.ca*

[d]*David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada*

[e]*Faculty of Engineering and Applied Science, Memorial University, St. John's, NL A1B 3X5, Canada*

[f]*Center for Catalysis and Separations, Khalifa University, P. O. Box 127788, Abu Dhabi, United Arab Emirates*

[g]*Department of Chemical and Petroleum Engineering, Khalifa University, 1274 Abu Dhabi, United Arab Emirates*

made recently to enhance the carbon capture and separation performance of MOFs.[12] Furthermore, some studies that evaluate the potential of these materials for use in $CO_2$ capture systems in the industrial chemical and energy sectors are emerging.[10]

For assessing and choosing adsorption methods, scientists identify the best combination of adsorbent structures for carbon capture processes. Discovering materials that have already been synthesized is one strategy, while another is to use simulation approaches to assess MOFs before synthesis.[11] The concept of developing an artificial intelligence (AI)-based chemical assistant has opened up previously unheard-of possibilities to revolutionize materials research, particularly for MOFs used in carbon capture. Today, time-consuming and difficult operations, such as data analysis, chemical screening, and library searches, can be processed quickly by utilizing AI expertise across multiple disciplines.[13] Employing a fixed vacuum swing adsorption (VSA) process, Burns *et al.*[14] evaluated over 1600 MOFs for post-combustion $CO_2$ capture by combining atomic/molecular simulation and process modeling. Only 500 of the above MOFs fulfilled the U.S. Department of Energy's standards for 90% purity and 95% recovery of $CO_2$. To facilitate this search, they created a machine learning (ML) algorithm that detects promising materials with an accuracy of over 90% and rapidly determines which materials should satisfy the purity and recovery requirements. Therefore, it is essential to employ AI/ML algorithms to accelerate the selection and optimization of the desired MOF structure for customization in $CO_2$ adsorption processes. One of the most significant challenges in chemistry and materials science research has been determining chemical compound information and material compositions, including optimal synthesis conditions as well as physical and chemical properties. A fundamental and crucial stage in the materials discovery process is to obtain a thorough summary of chemical information taken from literature sources, including articles and patents, and then store it in an orderly database.[15]

In general, scholars are particularly interested in effectively extracting vast volumes of material structure information from published scientific articles and existing literature. Natural language processing (NLP) models, which can quickly read and understand the words and information in published papers, are currently one of the most widely used methods in this field.[16,17] Large language models (LLMs), especially the GPT series, are emerging nowadays, and the fields of materials science and chemistry are undergoing a significant revolution because of these language models.[18]

One of the primary objectives is to use published data and screen them to extract required information and features for the design of MOF structures for various $CO_2$ adsorption processes. This valuable information can provide insights and promising opportunities for materials design researchers, seeking to synthesize and design a new generation of MOF compounds for $CO_2$ capture. Understanding the impact of each of the fundamental MOF properties, such as modifications to ligands, metals, and component functional groups, as well as the adsorption process conditions (temperature and pressure), is essential for the innovative design of MOFs for $CO_2$ separation from gases. The selection of the component structure or the synthesis method can be significantly influenced by the MOF synthesis techniques and the impact of each of the previously listed criteria.[19]

In the present study, we performed text data mining of scientific experimental data published in reputable journals on the synthesis of MOFs, specifically for $CO_2$ capture, using the LLM (ChatGPT) model. Compared to other generative AI platforms or other open-source models, we selected the ChatGPT 4o mini[20] because of its universal accessibility, computational correctness, task-based performance, and token limitations. First, one of the easiest generative AI systems to use is OpenAI's ChatGPT family. Unlike alternatives that require an application programming interface (API), interacting with a GPT model using a web user interface (UI) does not require any hardware or installation. Second, the GPT-4 series also demonstrates stronger performance in chemical science and engineering, including improved accuracy in tasks such as structure elucidation and reaction prediction.[21–23] These features create a new opportunity to use data analysis for accelerating research in this field. The findings from text mining on the synthesis methods of different MOFs for $CO_2$ capture were linked to process adsorption conditions and MOF performance. Finally, an analytical investigation was conducted to examine the relationships among synthesis techniques, structural components of MOFs, functional groups, and $CO_2$ capture performance. It provides a useful tool for decision-making and synthesis strategies, demonstrating that linking various MOF components with their corresponding synthesis techniques based on published papers is feasible. Furthermore, we used details collected from the synthesis conditions and types of MOFs for $CO_2$ adsorption to create a recommendation system for synthesis conditions. This method offers an asset for various MOF methods of synthesis by recommending customized MOF synthesis conditions based on specific metal and ligand types and a direct correlation with the amount of $CO_2$ adsorption. This research study demonstrates how LLMs can aid in chemistry and accelerate MOF customization to build high-efficiency $CO_2$ adsorbents.

## 2. Materials and methods

Finding accurate information from trustworthy sources is the main challenge; to overcome this issue, we need to select excellent, highly referenced articles for creating a reliable database in this area. Thus, we focused on papers, with the titles "Synthesis of MOFs for $CO_2$ capture," "MOF for $CO_2$," and "MOF for $CO_2$ adsorption," published in respectable journals. All keywords for searching articles with relevant topics, such as carbon capture using MOF, are mentioned in Section 1 (procedure section) of the SI with full details. We targeted 433 experimental articles from the vast collection of published articles in this field as input data, focusing on studies of MOFs and their composites that are employed only for $CO_2$ capture. The types of metals, functional groups, solvents, ligands, specific surface area (SSA), pore size, pore volume, reaction time, oxidation number, adsorption temperature and pressure, as well as $CO_2$

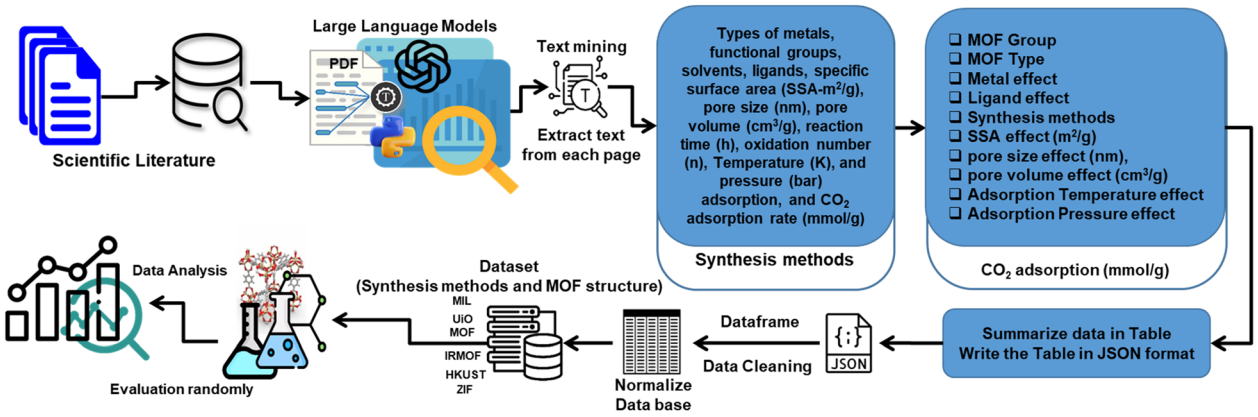


Fig. 1 General schematic of the LLM in the general process of extracting information and data from papers (PDF files), and the general process of text mining based on the defined features of the types of metals, functional groups, solvents, ligands, SSA ($m^2\ g^{-1}$), pore size (nm), pore volume ($cm^3\ g^{-1}$), reaction time (s), oxidation number (*n*), temperature (K) and pressure (bar) adsorption, $CO_2$ adsorption rate (mmol $g^{-1}$), MOF group, and synthesis methods. In this method, data integration is handled either directly by ChatGPT or through Python code written by ChatGPT. All prompts are performed by connecting the generated Python code to ChatGPT. The Python logo displayed is attributed to the Python Software Foundation, and the OpenAI logo is attributed to OpenAI.

adsorption rate are parameters that should be considered when selecting and organizing research articles to read and extract information on MOF synthesis conditions for $CO_2$ adsorption by ChatGPT 4o mini.[20] Furthermore, various writing styles need to be considered. In this study, we encountered a broad variety of writing styles that lacked a common framework across all published papers. Therefore, reading irrelevant information can be a challenge in identifying vital information about MOF compounds in scientific articles and published literature. Due to the various naming and spelling variations in scientific articles and literature, we adapted common keywords from these sources. Then, we had to identify all the complex and diverse information considered in the articles, such as MOF synthesis for $CO_2$ adsorption. We collected all the info/details on how to search and classify the articles, along with the selection of keywords in the SI file in the Methods section.

Fig. 1 presents a procedure for chemical text mining instructions, including how to extract all the information and name the MOF groups. The goal is to extract all data/information on MOF synthesis, including the name of compounds, metal source, ligand, solvent, and reaction time. Simultaneously, the $CO_2$ adsorption rate at different temperature and pressure conditions, along with pore size, pore volume, and SSA of MOFs, are other keywords. The details of the dataset used for chemical text mining are listed and compiled in the SI file. Fig. 1 is a schematic of the data mining workflow using ChatGPT for text mining and extracting information on MOF synthesis conditions and key parameters for $CO_2$ capture from a number of published research articles. In the first step, the articles' PDF files are considered as input data; during the initial evaluation, we take into account all the expected data displayed in the blue box that have an impact on $CO_2$ adsorption (see Fig. 1). The white or black OpenAI logo signifies the use of the ChatGPT platform, and the entire article review process is performed according to ChatGPT 4o-mini[20] language patterns. We created prompts for ChatGPT to guide its reading of the articles. By specifying keywords, the language model concentrates solely on the titles and terms we identified at the outset. This technique swiftly eliminates all parts unrelated to the article's keywords. By establishing this method, we enhance the processing and text-mining speed of our desired articles. In this approach, ChatGPT reads all selected keywords, along with the associated numbers and organizes them in a table as reference data (see Table S1). The texts of the articles are thoroughly scanned, and a table containing the MOF synthesis information, along with the $CO_2$ adsorption data and MOF specifications mentioned in the articles, is created. The data-gathering text mining procedure that employs ChatGPT4o-mini to collect information on MOF synthesis conditions and important $CO_2$ capture parameters from several published research articles is presented as a flowchart in Fig. S1 of the SI file. Instead of using individual, time-consuming chats with web-based ChatGPT to process text from numerous research papers, OpenAI's GPT-4o-mini, which is the same as the one that powers the ChatGPT product, allows for a more effective method because it has an API that allows text from a large number of papers to be processed in batches.

## 3. Results and discussion

### 3.1. Text mining performance

The ChatGPT extracts the desired data (the prompts mentioned in Fig. 2) from the articles PDF files using the OpenAI API. The extracted data is stored in a table. Since the data are compiled from different articles, they are not uniform in chemical compound names and units. In addition, the data is sometimes stored in a random combination of numbers and letters in the table cells. In fact, ChatGPT's output table for data analysis is very messy. Therefore, to prepare and sort the data for subsequent analyses, an extensive process was carried out to standardize the data table. During preprocessing, extensive modifications were made to the data. These modifications consist of four steps. (i) For correlation between the data, first,

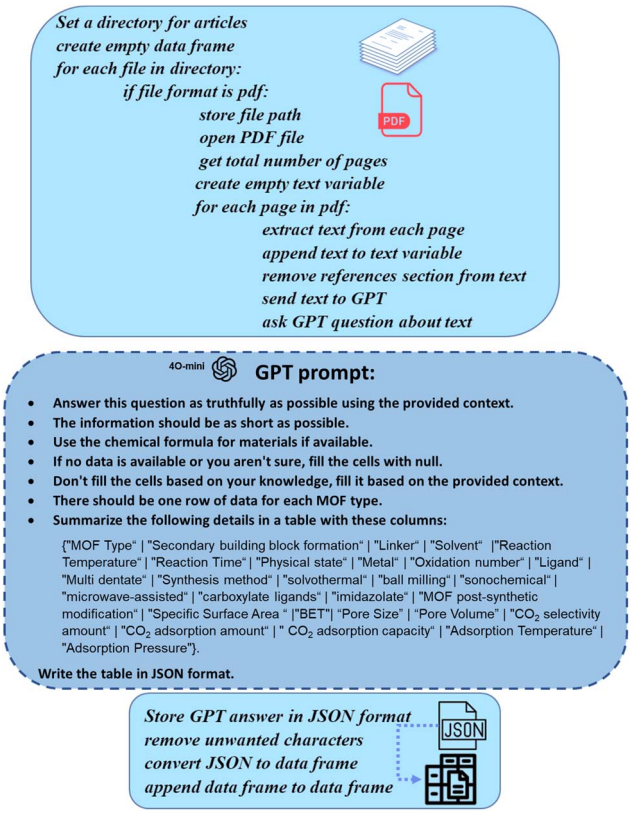


Fig. 2 Pseudo-code of the GPT-4o mini prompts with the LLM algorithm interaction for the MOF parameters.

unusual data were removed from the table as a whole, including duplicate rows and rows with insufficient data. (ii) Extensive unit conversion was performed. Since the articles reported numbers in various units, all numerical data were standardized to the same unit. (iii) Non-numeric data, such as chemical names of ligands and solvents, were standardized. (iv) Scattered and diverse names were placed in larger and more comprehensive classifications. All these stages were carried out to clean and standardize the data.

Fig. 2 shows the pseudo-code and addresses the corresponding effects of these parameters using an LLM-based model developed to establish a meaningful relationship between the parameters. Given the limited availability of $CO_2$ selectivity data, it was not possible to find a robust relationship between this parameter and the other parameters. All statistical data for pressure, $CO_2$ adsorption temperature, SSA, and oxidation number shown in Fig. S8 to S10 in the SI have been reported in studies that focused on adsorption pressure, adsorption temperature, reaction time, and oxidation number. The data cover oxidation numbers from 0 to 4, temperatures up to 325 K, and adsorption pressures up to 15 bar. Fig. S7 in the SI document shows the scatter of the data across the top ten ligands, metals, and MOF groups.

As shown in Fig. 1, textual data about different MOFs, such as MOF types, ligands, metals, synthesis method, solvent, physical states, oxidation number, reaction time, MOF structure characteristics, SSA ($m^2$ $g^{-1}$), pore size diameter (nm), pore volume ($cm^3$ $g^{-1}$), $CO_2$ adsorption capacity (mmol $g^{-1}$), $CO_2$ selectivity, adsorption temperature (K), adsorption pressure (bar), and MOF group name, which are present in the SI, were extracted as the data required for implementation of LLMs. Comprehensive details about MOF group characteristics, synthesis methods, and $CO_2$ adsorption rates are provided in Table S1 of the SI. By carefully extracting the data and applying one-hot coding to the MOF group names and related features, appropriate weighting was performed. The results in Fig. S4 reveal the relationship between the mentioned parameters in the SI. Pearson correlations between all the considered parameters for the top five metals, along with the top six MOFs, structural variables, and $CO_2$ adsorption characteristics, are plotted in Fig. S4 of the SI.

All MOF structural characteristics, including the SSA ($m^2$ $g^{-1}$), pore size diameter (nm), pore volume ($cm^3$ $g^{-1}$), and $CO_2$ adsorption capacity (mmol $g^{-1}$), of the top five metals, and ligands, which were obtained from the data text mining approach, are shown in Fig. 3–6. The top six ligands (H3BTC, H2BDC, H4DOBPDC, 2-methylimidazole, H3BTB, and $NH_2$-BDC), metals (Cu, Zn, Mg, Cr, Zr, and Co), MOF groups (UiO-66, MIL-101, MIL-100, HKUST-1, MOF-5, and MOF-74), and top five supergroups (MOF, MIL, ZIF, UiO, and metal/dobpdc) are listed in (Fig. 4, 6, 8 and 9). This means that the statistical data on the

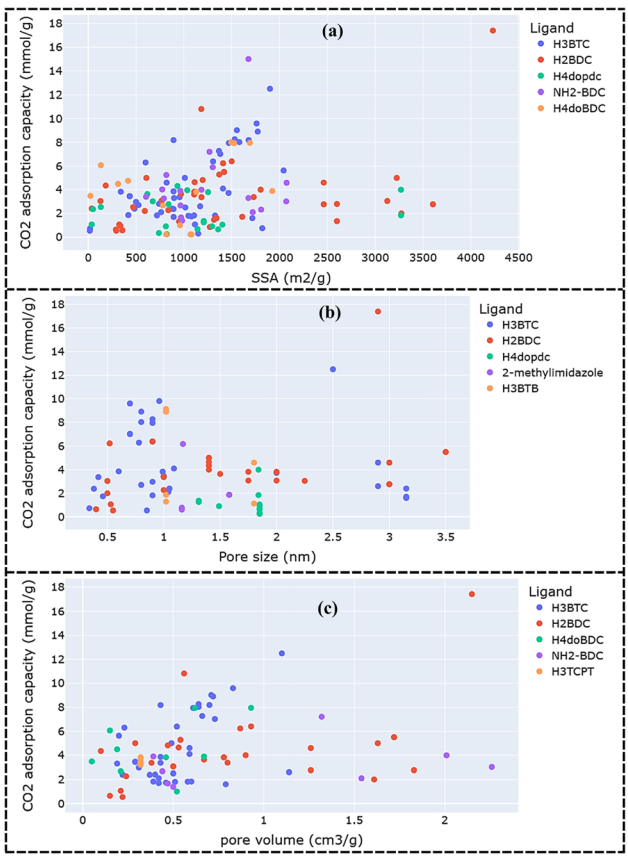


Fig. 3 (a) $CO_2$ adsorption capacity (mmol $g^{-1}$) data range for the top five ligands, along with the SSA ($m^2$ $g^{-1}$), (b) pore size (nm), and (c) pore volume ($cm^3$ $g^{-1}$), obtained for selected MOF.

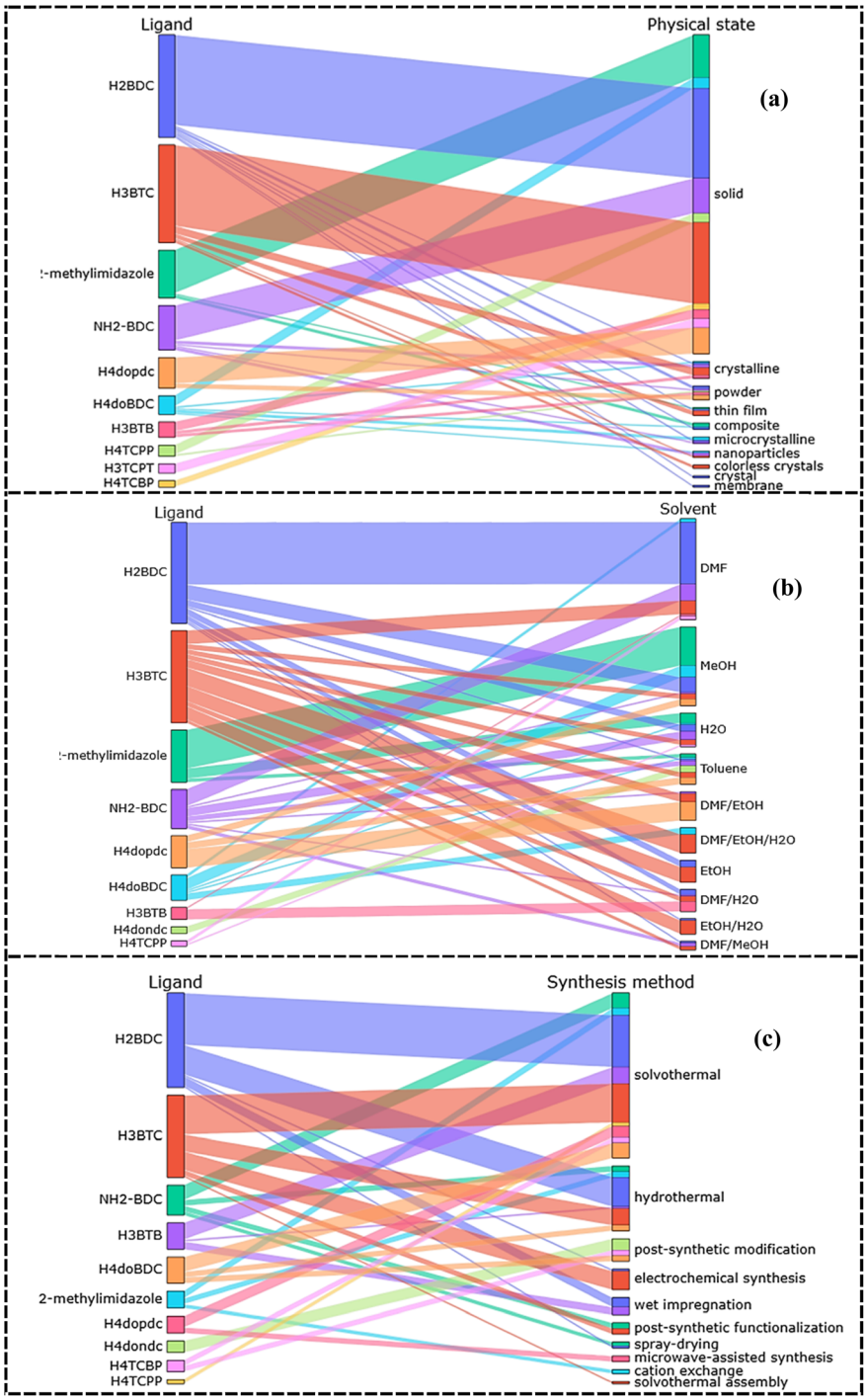


Fig. 4 Relationship between the $CO_2$ adsorption capacity data range in Fig. 3 for the top ten ligand types introduced for (a) ten physical states, (b) ten solvents, and (c) ten MOF synthesis methods.

top parameters had more repeatability and more data. The execution method and instructions for these prompts are presented in Fig. 2. To understand how the GPT-4o mini[20] prompt interacts with the LLM algorithm, the pseudo-codes are demonstrated in Fig. 2 and the S1 document. The data extraction process began by analysing individual paragraphs in each article, after first separating the articles into independent PDF files. To coordinate interactions between the different LLMs, we used the OpenAI API to perform data mining on the documents. To read the PDF documents and convert them into text, PyPDF2 version 3.0.1 was utilized. All data manipulation, analysis, and processing were performed by employing Panda's library version 1.4.3. Diagrams and figures were plotted by the Plotly library version 5.13.1. The workflow enables fully automated data extraction, which includes three main steps: namely classification, feature consideration and inclusion, and finally

 

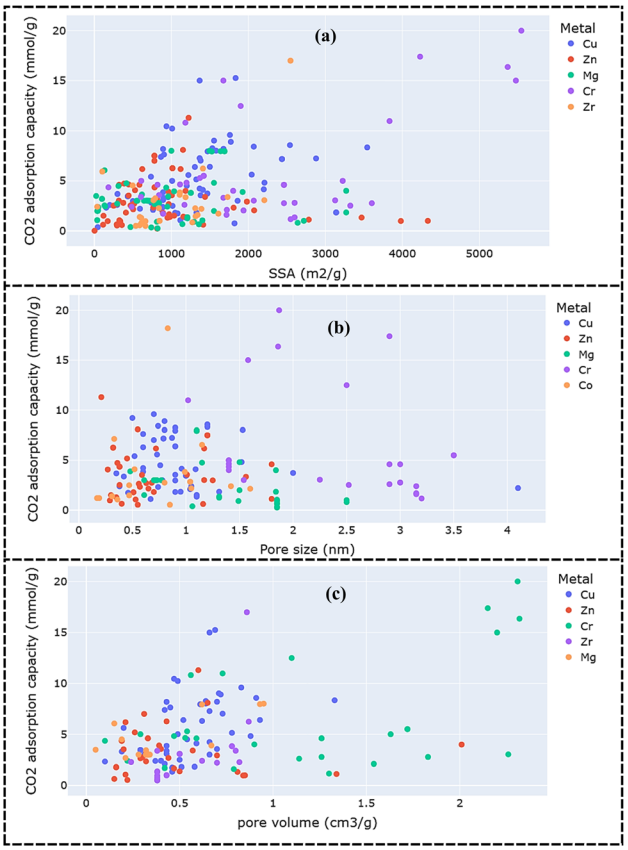


Fig. 5 $CO_2$ adsorption capacity (mmol $g^{-1}$) data range for the top five metals *versus* (a) SSA ($m^2$ $g^{-1}$), (b) pore size (nm), and (c) pore volume ($cm^3$ $g^{-1}$), for considered MOF.

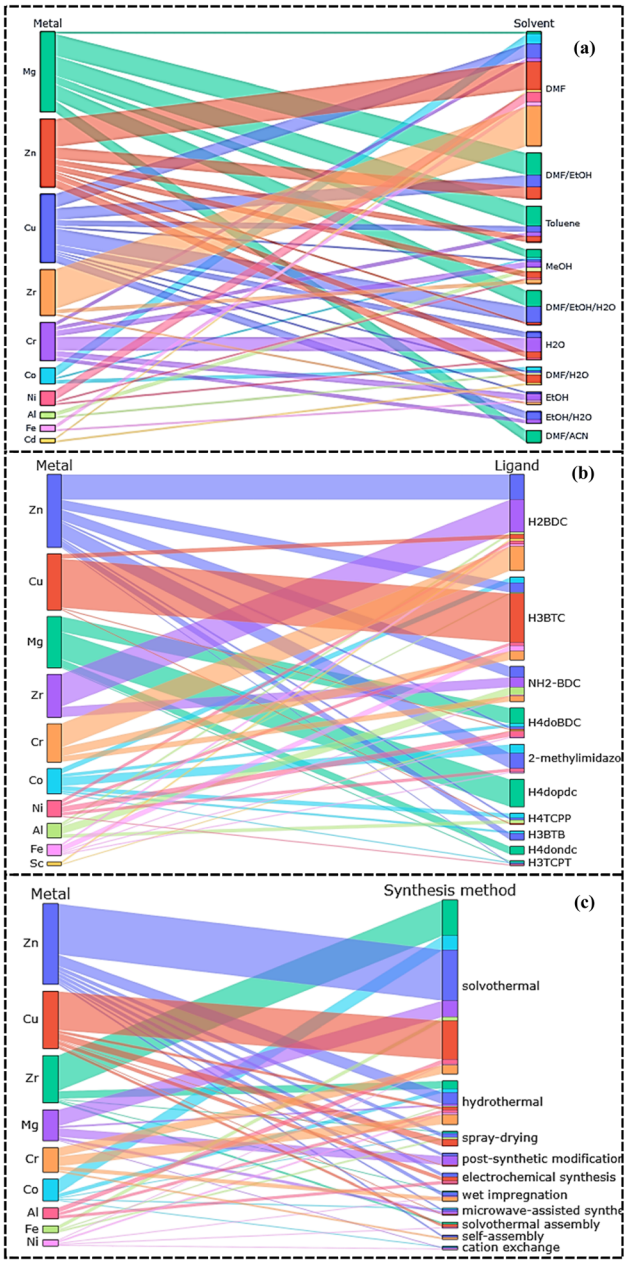


Fig. 6 Relationship between the $CO_2$ adsorption capacity data ranges in Fig. 5 for the top ten metal types introduced for (a) ten solvents, (b) ten ligands, and (c) ten MOF synthesis methods.

extraction. The automation was performed manually with modifications such as monitoring token length. Table S2 illustrates the accuracy achieved for the classification and inclusion of features. Fig. S1 in the SI document depicts the data mining path classification and data cleaning procedure involved in this process. In Table S2, where the evaluation results are specified, the column "Description by human reviewer" lists the observed discrepancies, and the results of the review of the articles that had discrepancies are listed in the column "Status".

After extracting the identified data and establishing the appropriate data format, a validation process was performed by comparing the extracted data with the original articles to minimize errors in 10% of the dataset (see SI Excel, Table S2). The text and keywords were correctly detected in over 86% of the ChatGPT readings. To check the accuracy of the data extracted by ChatGPT, 10% of the articles were randomly selected (using the random function) and were reviewed manually and by a human reviewer. The results of this review are given in Table S2 in the SI. We decided to remove $CO_2$ selectivity from the results in the table because it did not show a significant relationship with the other parameters. Furthermore, as indicated in Table S2, inconsistencies in how selectivity was reported/written across studies (combined with the limited amount of available data) made meaningful analysis difficult. Thus, this parameter was excluded from further consideration.

Fig. S4(a) highlights the key correlations between $CO_2$ adsorption parameters, metals, pore volume, specific surface area (SSA), and adsorption pressure with other parameters. Fig. S4(b) shows that SSA and pore volume strongly influence the adsorption rate of metals Cr and Cu, while adsorption temperature exhibits the strongest correlation with Cr and Mg. Fig. S4(c) displays the significant structure-property relationships, where MIL-101 shows the strongest SSA–SSA-adsorption correlation, MOF-74 demonstrates the highest adsorption capacity correlation, and MIL-100 shows the strongest temperature correlation.

We should focus on reducing the illusion created by the comparison table between real data extraction and the fabricated or misleading content from ChatGPT. Examples of what is or is not in the text (Table S1) and how ChatGPT presents it in a fabricated state can be seen in the SI Excel file (Table S2). Therefore, the issue of illusion is an important concern, and we need to analyze and evaluate the obtained data so we can control how we retrieve data while designing ChatGPT commands. The $CO_2$ adsorption rates (mmol $g^{-1}$) with related MOF structure parameters such as SSA ($m^2$ $g^{-1}$), pore size (nm), and pore volume ($cm^3$ $g^{-1}$) using the top five ligands listed in the literature are illustrated in Fig. 3(a–c). After investigating the data, ref. 24–31 are consulted to verify the accuracy of these items/factors. The findings indicate that the H2BDC ligand has the highest porosity, and the H3BTC ligand has the greatest $CO_2$ adsorption capability. Fig. 4(a–c) displays the relationship diagram of the top ten ligands in terms of physical state, solvent, and synthesis method parameters, respectively. The connections between each ligand and the synthesis parameters are evident in these figures. The findings offer researchers a broad framework for synthesis and can serve as useful guidelines when using different MOF synthesis techniques for $CO_2$ adsorption application. The amount of $CO_2$ adsorption, along with the corresponding SSA, pore size and pore volume for particular MOF ligands (frequently reported in the literature), and their synthesis methods, serves as evidence of the reproducibility of the information obtained. This indicates that data gathered from the literature can be used for synthesizing MOFs targeted for $CO_2$ capture.

The hexanuclear $[Zr_6O_4(OH)_4]$ units that make up the hydroxylated form of UiO-66 have $\mu_3$-O and $\mu_3$-OH groups alternately capped on the triangular faces of the $Zr_6$ octahedron. To create a cubic 3-D framework, the $Zr_6$ polyhedra are joined along their edges by carboxylate groups from twelve 1,4-benzenedicarboxylate (BDC) linkers.[32] Experimental and simulation studies have demonstrated that the addition of functional groups like –COOH, $–SO_3H$, and $–NH_2$ to the BDC ligand significantly enhances UiO-66's capacity to adsorb and separate $CO_2$.[33] Since UiO-66 structures often have smaller pore diameters and surface areas, adding functional groups generally reduces the adsorbent's SSA, which may harm the $CO_2$ adsorption capacity of porous adsorbents.[34]

The same strategy is repeated in Fig. 5–9 for the top five metals and MOF groups identified, along with their relationship to physical conditions, solvents, metals, ligands, and synthesis methods. The relationships between the top five metals and synthesis parameters, along with the $CO_2$ adsorption rate and its relationship to SSA, pore size, and pore volume, are evident in these figures. Using these findings, the identification of the top five metals and their interactions with the $CO_2$ adsorption rate (mmol $g^{-1}$) results and the structure of the MOF in terms of porosity, SSA, and pore size can be understood. By combining this information with the various methods of MOF synthesis, researchers can make more informed decisions regarding MOF selection, targeted adsorption performance, and desired porosity characteristics. Organizing and categorizing these data together provides a clearer research roadmap and makes potential objectives more achievable. For example, Cu metal created the porous structure with the highest SSA ($m^2$ $g^{-1}$) and $CO_2$ adsorption capacity (mmol $g^{-1}$) among all metals. Fig. 4, 6, and 8 illustrate another aspect of the top five metal interactions with ligands and MOF groups, as well as synthesis details, including the solvent type and synthesis methods.

The sources involving UiO-66 and MIL-101 were investigated further, and the text mining findings were in agreement with every result reported in the relevant papers.[35,36]

Fig. 3–6, concerning the top five ligands and metals, confirm the results of Fig. S11 (in the SI file) and the relationship between the aforementioned parameters. This approach allows for flexibility in selecting different synthesis conditions among the top ten ligands and metals. Furthermore, by providing an overview of the synthesis methods and the rationale behind solvent selection, researchers are better equipped to anticipate and understand the study approach. The first step is to establish a relationship between all synthesis factors, such as the choice of solvent, physical state, ligand, metal, and synthesis technique, with the $CO_2$ adsorption rate and porosity of the final MOF type. The next step is to observe and classify the MOF types and their grouping. An overview of these relationships is clearly illustrated in Fig. 3–8. Based on the statistical data from published articles, Fig. 7 and 8 further identify MOF-5, MIL-101, UiO-66, and HKUST-1 as the most effective MOF groups for $CO_2$ adsorption. The selection of a specific MOF group provides a framework for researchers to investigate the effects of key variables, such as the metal type, ligand, synthesis method, and solvent used, on overall $CO_2$ capture performance. An evaluation of room-temperature adsorption diffraction data implies that a secondary adsorption site contributes to the adsorption behavior of many of these materials. Fig. 8 and 9 show the relationship between ligands, metals, solvents, synthesis methods, and MOF type. Indeed, the collected data may be easily linked to the experimental results, serving as a useful reference for future study. In all Fig. 4, 6, 8, 9a, and 9b, direct relationships between top metal groups, top ligands, top solvents, top synthesis methods, and top MOF groups that had the greatest effect on $CO_2$ adsorption are shown.

It is worth noting that the roles of each metal and ligand in relation to the synthesis methods can be observed. For example, Cu demonstrates a progressively stronger influence in MOFs as the pore volume varies. The synthesis routes, extracted from text mining and illustrated in Fig. 6–9, provide further insight into these relationships. In addition, Fig. S8 and S9 depict statistical data distributions of other key parameters influencing $CO_2$ adsorption capacity, including the top five ligands, metals, and MOF groups. The results confirm the impact of the top five ligands as well as the top five metals on $CO_2$ adsorption. Cu and Co metals, along with $H_2BDC$ and $NH_2BDC$ ligands from MOF-74 and MIL-101, will lead to a high specific surface area and enhanced $CO_2$ adsorption (see Fig. S11). Therefore, based on data obtained using solvothermal, hydrothermal, spray drying, and post-synthetic modification methods and utilizing DMF, MeOH, and $H_2O$ solvents in the presence of $H_2BDC/NH_2$-BDC ligands and Cu, it is possible to achieve MOF composites with greater $CO_2$ adsorption capacity. The results reveal that text

 

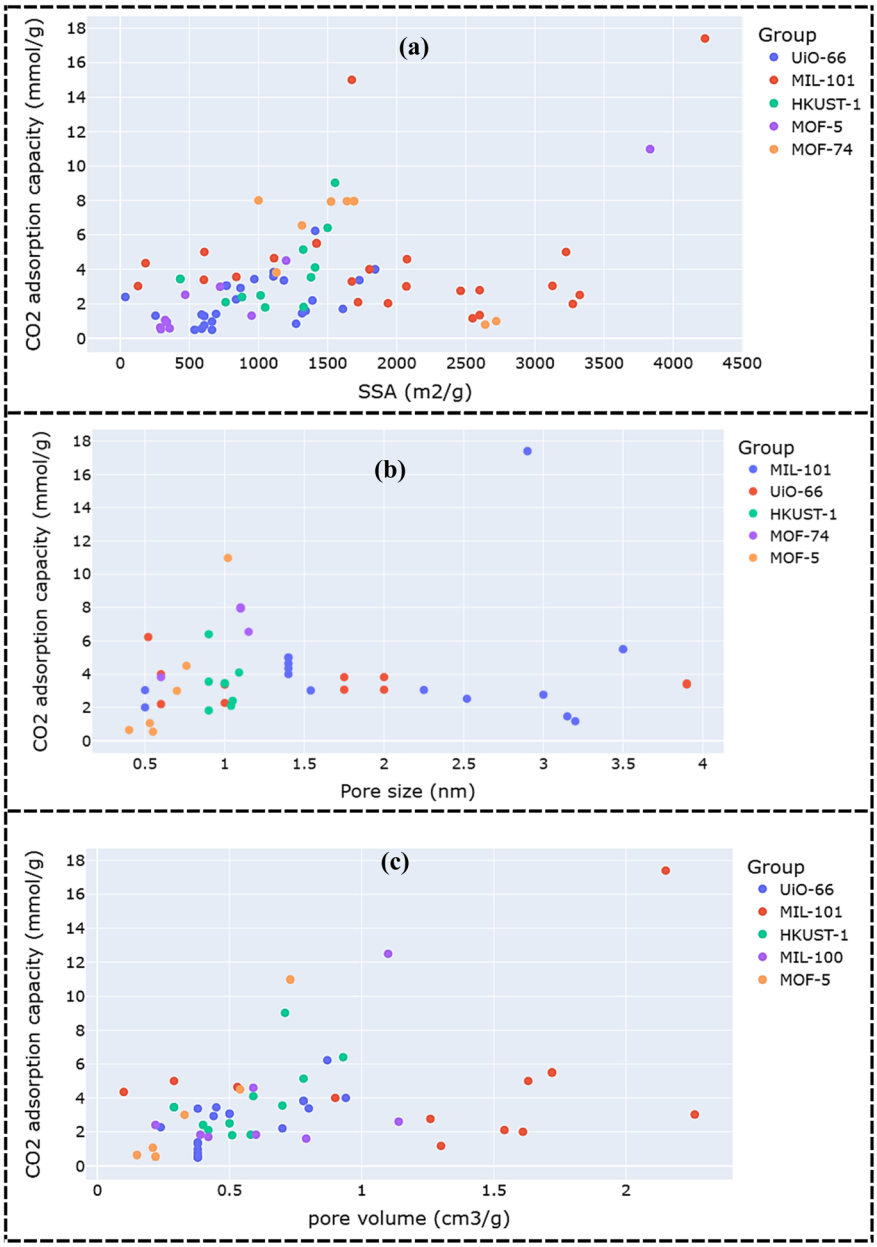


Fig. 7 $CO_2$ adsorption capacity (mmol $g^{-1}$) data range for the top five MOF groups, in terms of (a) SSA ($m^2$ $g^{-1}$), (b) pore size (nm), and (c) pore volume ($cm^3$ $g^{-1}$), obtained for considered MOF in Fig. 8.

mining facilitates data-driven decision-making to optimize the desired MOFs for $CO_2$ uptake or synthesis method selection. By empowering researchers to choose synthesis methods, solvents, ligand types, and metals, and to understand how these factors interact to shape the MOF structure and performance, we move closer to developing a smart assistant. This assistant, supported by text mining, will enhance researchers' ability to select the best synthesis techniques, solvents, ligand types, and metals, as well as to understand how these features interact to create the optimum MOF structure with high performance.

As shown in Fig. 3–9, the textual data analysis highlights the influence of different metals (*e.g.*, Cu and Zn) and ligands (*e.g.*, BDC) on adsorption efficiency. Variations in metal centers and ligand types significantly affect the porosity of the adsorbent and the resulting MOF structure. Recent studies indicate that Zn-based MOFs, particularly those incorporating $Cu_2BDC_2$ frameworks, exhibit strong and rapid $CO_2$ uptake.[37] Furthermore, the effect of pressure on MOF-based adsorbents, as extracted from the textual data, provides valuable guidance for process optimization before experimental synthesis. As an example of validating our results against published studies, we assessed the MOF-74 sample and compared the data obtained through text data mining with atomic simulation data and experimental data for transition-metal MOF-74 variants. It is noticed that the results of text mining, which are based on experimental data, would support the findings of this study (see

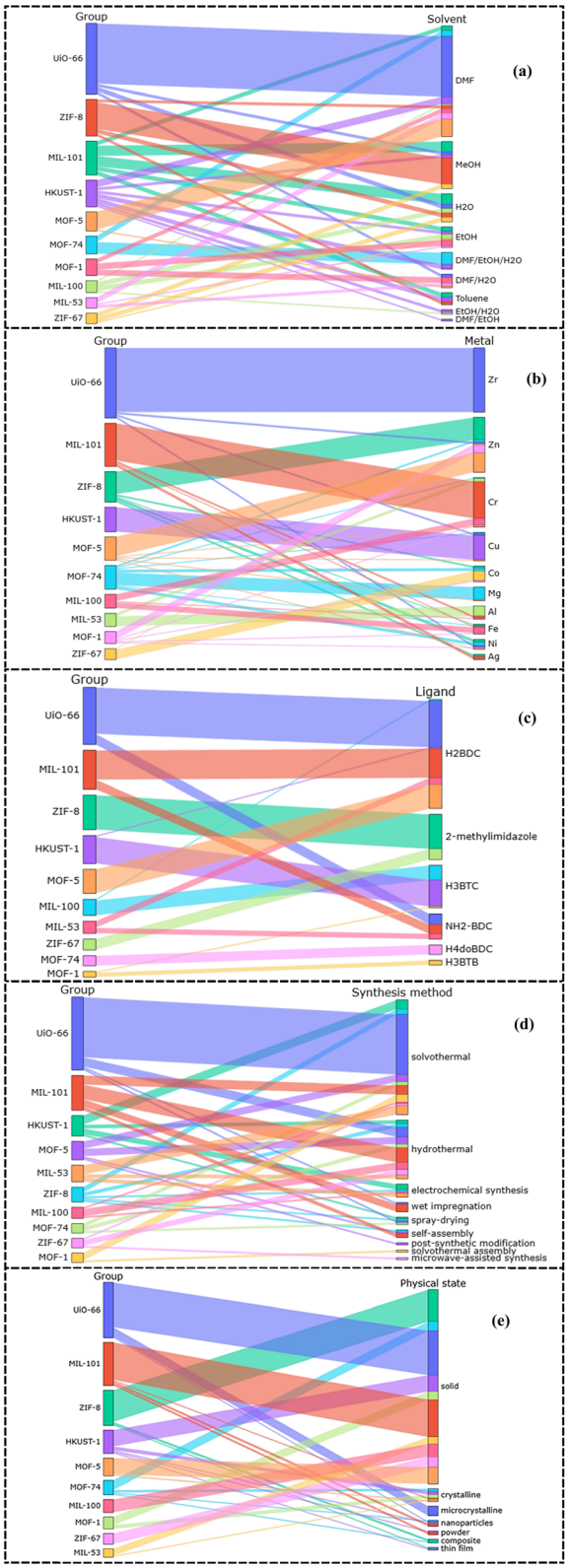


Fig. 8 Relationship between the $CO_2$ adsorption capacity data ranges in Fig. 7 for the top ten MOF group types introduced for (a) nine solvents, (b) ten metals, (c) six ligands, (d) nine synthesis methods, and (e) seven physical states.

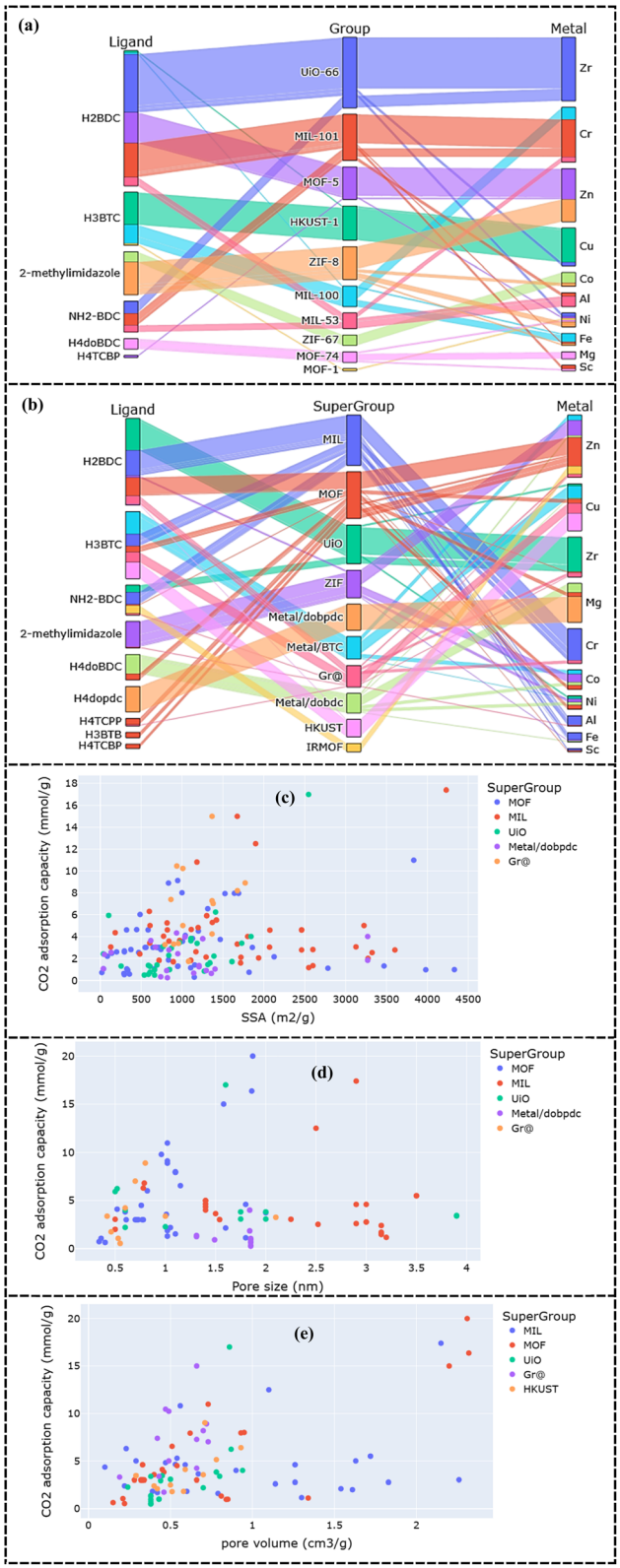


Fig. 9 (a) Relationships and classification of MOF groups and (b) supergroups with metals and ligands obtained from text mining. Evaluation of the top five MOF supergroups based on $CO_2$ adsorption performance (mmol $g^{-1}$) in different ranges, along with the MOF features of (c) SSA ($m^2$ $g^{-1}$), (d) pore size (nm), and (e) pore volume ($cm^3$ $g^{-1}$), extracted from text mining.

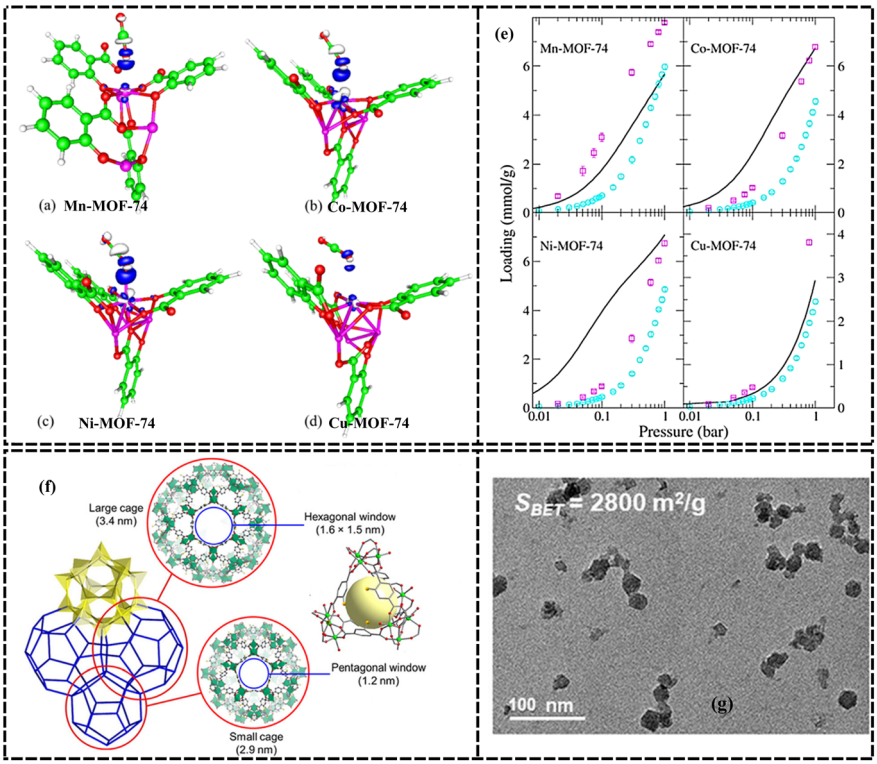


Fig. 10 Density distribution plots of the metals of the MOF-74: (a) Mn-MOF-74, (b) Co-MOF-74, (c) Ni-MOF-74, and (d) Cu-MOF-74 model clusters interacting with a $CO_2$ molecule, and (e) $CO_2$ adsorption isotherms for Mn, Co, Ni, and Cu-MOF-74, utilizing universal force field (UFF) in conjunction with the localization of properties and density-derived electrostatic and chemical (cyan circles) point charges, calculated at a temperature of 298 K. Reprinted with permission from ref. 40 Copyright 2015, the American Chemical Society. Experimental isotherms at 298 K are displayed for comparison (black curves, data extracted from ref. 38 Copyright 2014 with permission from the Royal Society of Chemistry). (f) Schematic of the MIL-101(Cr)-$NH_2$ structure[39,41] and (g) TEM images of MIL-101(Cr)-$NH_2$ nanoparticles. Reprinted with permission from ref. 39 Copyright 2020, the American Chemical Society.

Fig. 10). Queen *et al.*[38] found that the Cu analogue has an axial strain that causes the ligand $O_2$ atom to be positioned in such a way that $CO_2$ cannot approach the open metal site. Since the $CO_2$ site's occupancy is similar, it is clear that $Cu_2$(dobdc) possesses two adsorption sites with identical binding strengths. The results obtained from text data mining are in strong agreement with the experimental results[39] of the role of Cr and ligand for tuning the pore size and SSA for the synthesis of amine-functionalized MIL-101(Cr)-$NH_2$ with a particle size less than 20 nm, SSA above 2800 $m^2\ g^{-1}$, and $CO_2$ adsorption up to 3.4 mmol $g^{-1}$ (see Fig. 10 (f) and (g)).

Our strategy, with the help of the ChatGPT-4o-mini LLM assistant, aims to quickly review a variety of synthesis methods, including green methods, to facilitate routine experimental work. Using the extensive knowledge previously published by researchers, we enable rapid searching and evaluation of various synthesis parameters by utilizing AI algorithms. The standard criteria of gas adsorption by porous materials depend on various parameters, such as (i) variations in size and/or shape (molecular sieve effect); (ii) variations in the interactions between the adsorbent molecule and adsorbent surface (thermodynamic equilibrium effect); (iii) variations in diffusion intensities (kinetic effect or partial molecular sieve action); and (iv) quantum effects.[42] The interaction between the gas and the MOF surface becomes a key parameter in determining the quantity of adsorption of each component when the MOF adsorbent's pore size is large enough to allow all gas components to pass through. Additionally, the characteristics of the adsorbent, such as polarizability, permanent dipole moment, quadrupole moment, as well as the features of the adsorbent surface, influence the interaction intensity.[43] When carboxylate ligands (1,4-benzenedicarboxylates (BDC)) are combined with high-valence metal cations ($Cr^{3+}$), the MOF stability is improved in the presence of water. However, $CO_2$ adsorption, which produces a quadrupole moment, is significantly affected by strong polarizing groups such as carboxylic acid.[44] Furthermore, increasing metal valence causes an increase in the electrical difference between the $CO_2$ molecule and the adsorbent surface, which improves $CO_2$ adsorption. For $CO_2$ adsorption, the adsorbent's pore diameter (3.3 Å) must be greater than the kinetic diameter of $CO_2$ molecule. Because of the different interactions between the adsorbent and its surface, $CO_2$ gas is adsorbed, and the adsorption process will be at thermodynamic equilibrium.[12] Consequently, it is feasible to anticipate increased $CO_2$ adsorption and tune the pore size diameter of MOFs, by carefully selecting the appropriate ligands and metals.

Fig. 11 provides an overview of the creation of keywords and classification of the MOF group for $CO_2$ capture. Based on the

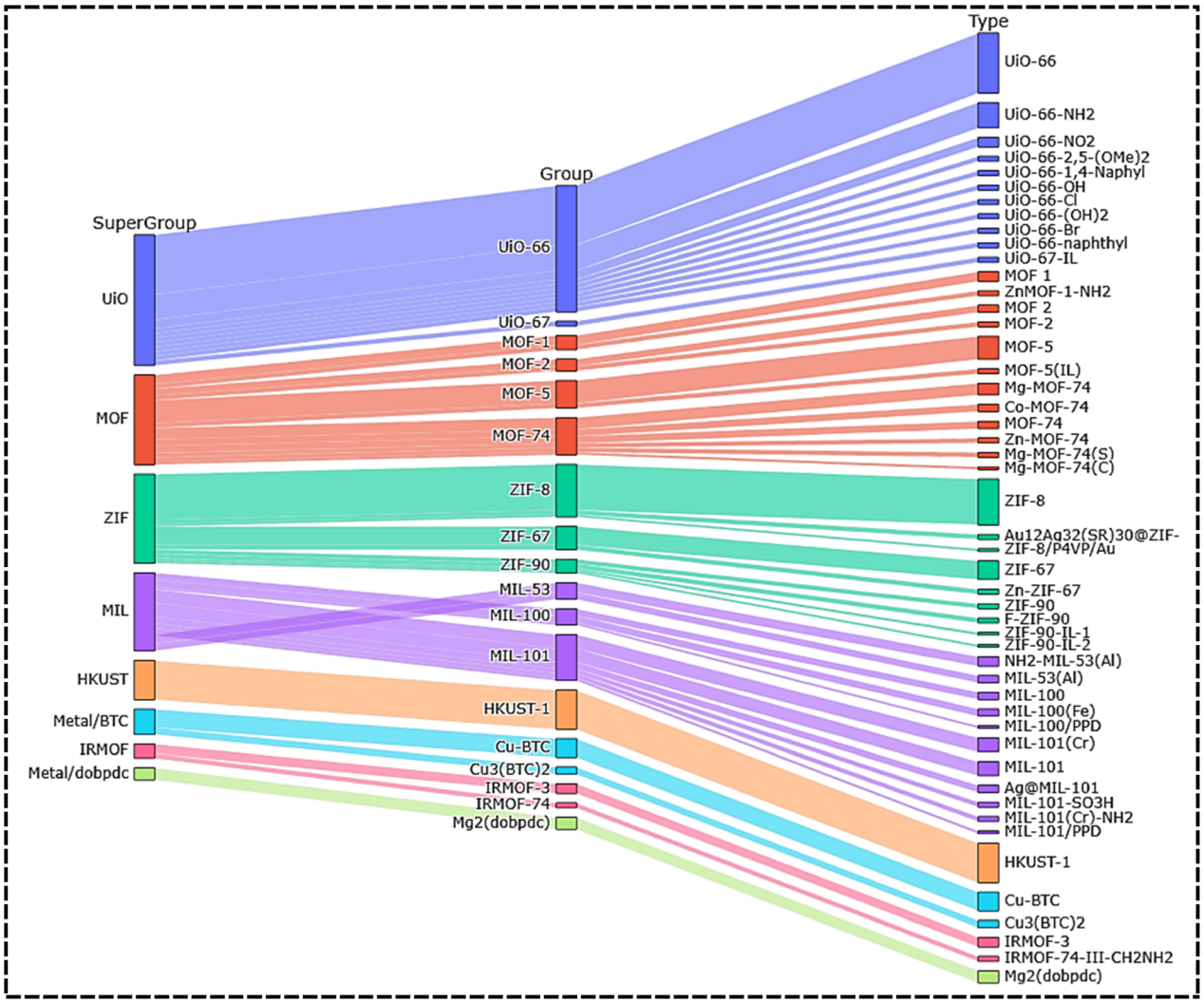


Fig. 11 Schematic of the classified types of supergroups, MOF groups, and MOF types obtained from text mining.

classification generated by the LLM, MOFs can be organized into groups and supergroups, and the relationships among the MOF types used for $CO_2$ capture are shown in Fig. 9. In fact, the relationship of the ligand and metal types is one of the most important foundations of the MOF structure. By observing the type of MOF in the separated subgroups, the results of the mentioned diagrams can be easily used. Using the current ChatGPT-based assistant, researchers can gain an overview and a roadmap of how different MOF or composite types are synthesized and perform. This allows them to analyze the data effectively and make informed decisions when selecting an MOF type or synthesis method. New data can be integrated into the existing dataset to update the model and results using the attached GitHub code. The approach is easily expandable by incorporating additional data and refining the outputs accordingly. One of the main motivations behind this work was to promote the green synthesis of new MOFs by avoiding the use of toxic solvents such as DMF and reducing synthesis costs.[45] The resulting MOF cleaned database (Table S3, Excel file in the SI), including DOI links to the original articles, can be used asdata for training ML models and for evaluating various features.

## 4. Conclusions and outlook

In this research work, we investigated how the use of LLMs through the ChatGPT-4o mini platform can enhance the understanding and effectiveness of article text mining to assist chemists and chemical engineers in the design, synthesis, and utilization of MOF materials for $CO_2$ capture. We introduced a scheme for data mining and subsequent analysis of MOF data related to $CO_2$ capture using LLM, based on published articles. Our evaluation of the results and the data mining process demonstrated the potential of LLMs as an efficient tool for rapidly classifying and extracting MOF-related $CO_2$ capture data from published scientific articles. Statistical analysis of the data mining results from the articles provided extensive insights into the process conditions, porous structures of different MOFs, and synthesis methods relevant to $CO_2$ capture. Statistical differences were evaluated for 10% of the results, and data inconsistencies were assessed accordingly.

Based on human analysis, a significant percentage of correct data from the articles was identified and extracted by ChatGPT with appropriate accuracy and presented an acceptable true-

positive ratio. A major problem with the data in the articles is that ChatGPT fails to identify many instances of false negatives, particularly in review articles, where it only extracts a small number of relevant data items from a vast amount of information. This issue with $CO_2$ selectivity stopped us from adding this parameter to the database. Another issue with ChatGPT is its ability to identify data in large tables, as it extracts a limited number of items from them. Therefore, the results show that research article data is more practical and it is easier to read the text correctly by ChatGPT. Incorporating more research articles to obtain new data and creating a larger database will facilitate broader and more accurate research in materials science, especially on MOFs for certain applications. To improve the quality of LLM-based models, several key challenges need to be addressed. First, the literature is biased towards successful results, while unsuccessful experiments are rarely reported, which introduces uncertainty in the results. Second, text mining with LLMs is still in its early stages, and as the scientific literature continues to evolve, LLM results will change, potentially leading to inconsistencies. Therefore, it is imperative to address this concern and develop more linguistic models on specific topics, such as the synthesis of porous materials, with further validation. Therefore, one should not immediately expect accurate results from LLMs. Instead, they should be used with caution as an assistant to observation and data mining, material design, and laboratory synthesis to simplify and accelerate the review of past research. To improve validation, we recommend that researchers include the obtained data in a separate table in note format when publishing research results to allow for careful review and secondary validation. This research study not only demonstrates how LLMs can revolutionize the development of porous MOF materials for $CO_2$ capture, but also provides a useful guide for the design of high-performance MOF materials in the more general areas of designing and fabricating adsorbents and catalysts with efficient synthesis schemes. This practical tool can play a fundamental role for researchers in synthetic strategies and in providing roadmaps. Expanding this research method beyond adsorbent and catalyst studies would be encouraging. Crystal-structured adsorbents, such as zeolites and metal oxides, have high potential for future research.

## Nomenclatures

| | |
|---|---|
| $N$ | Number of datasets for training [-] |
| $R^2$ | Correlation coefficient [%] |
| $S_{BET}$ | Specific surface area [$m^2$ $g^{-1}$] |

### Acronyms

| | |
|---|---|
| AI | Artificial intelligence |
| ANN | Artificial neural network |
| AARD | Average absolute relative deviation [%] |
| AAD | Average absolute deviation [%] |
| API | Application programming interface |
| CCS | Carbon capture and sequestration |
| CBM | Carbon-based materials |
| GHG | Greenhouse gas |
| IPCC | Intergovernmental panel on climate change |
| MSE | Mean square error |
| MOF | Metal–organic framework |
| ML | Machine learning |
| MLP | Multi-layer perceptron |
| POP | Porous organic polymers |
| RMSE | Root means square error |
| SSA | Specific surface area [$m^2$ $g^{-1}$] |
| TEM | Transmission electron microscopy |
| VSA | Vacuum swing adsorption |

## Conflicts of interest

The authors state that none of their known financial conflicts or relationships could have had an impact on the work presented in this paper.

## Data availability

The codes, procedures and workflow used in this study are available in the GitHub repository, accessible *via* the link below.

https://github.com/ai4mat-lab/GPT_MOF_Project

Further details on usage can be found in the repository's documentation. A supplementary information (SI) data repository is available on Zenodo (https://doi.org/10.5281/zenodo.17619285), providing the datasets and codes used in this research study.

Supplementary information: detailed information about the dataset is summarized in Tables S2 (evaluated GPT results), and S3 (MOF database cleaned SI (Excel files)). Table S1 displays all extracted data from the PDF files. The results of the validation process between the ChatGPT text mining and comparison with the original article's experimental data are presented in Table S2. The dataset, which includes the exact paper's DOI and parameter values, can be found in Table S3. See DOI: https://doi.org/10.1039/d5dd00446b.

## Acknowledgements

This work is supported by the Natural Resources Canada through the Program of Energy Research and Development (PERD).

## References

1 H. Lee, K. Calvin, D. Dasgupta, G. Krinner, A. Mukherji, P. Thorne, C. Trisos, J. Romero, P. Aldunce and K. Barrett, *Climate change 2023: Synthesis Report. Contribution of Working Groups I, II and III to The Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, The Australian National University, 2023.

2 G. Luderer, Z. Vrontisi, C. Bertram, O. Y. Edelenbosch, R. C. Pietzcker, J. Rogelj, H. S. De Boer, L. Drouet, J. Emmerling and O. Fricko, *Nat. Clim. Change*, 2018, **8**, 626–633.

3 Y.-M. Wei, J.-N. Kang, L.-C. Liu, Q. Li, P.-T. Wang, J.-J. Hou, Q.-M. Liang, H. Liao, S.-F. Huang and B. Yu, *Nat. Clim. Change*, 2021, **11**, 112–118.
4 M. Pardakhti, T. Jafari, Z. Tobin, B. Dutta, E. Moharreri, N. S. Shemshaki, S. Suib and R. Srivastava, *ACS Appl. Mater. Interfaces*, 2019, **11**, 34533–34559.
5 M. T. Dunstan, F. Donat, A. H. Bork, C. P. Grey and C. R. Müller, *Chem. Rev.*, 2021, **121**, 12681–12745.
6 C. Zhang, R. Kong, X. Wang, Y. Xu, F. Wang, W. Ren, Y. Wang, F. Su and J.-X. Jiang, *Carbon N. Y.*, 2017, **114**, 608–618.
7 Z. Tao, Y. Tian, W. Wu, Z. Liu, W. Fu, C.-W. Kung and J. Shang, *npj Mater. Sustain.*, 2024, **2**, 20.
8 H. Mashhadimoslem, P. Karimi, M. A. Abdol, K. Zanganeh, A. Shafeen, A. A. AlHammadi and A. Elkamel, *Ind. Eng. Chem. Res.*, 2024, **63**, 11018–11029.
9 H. Mashhadimoslem, M. A. Abdol, K. Zanganeh, A. Shafeen, A. A. AlHammadi, M. Kamkar and A. Elkamel, *ACS Appl. Energy Mater.*, 2024, **7**, 8596–8609.
10 C. A. Trickett, A. Helal, B. A. Al-Maythalony, Z. H. Yamani, K. E. Cordova and O. M. Yaghi, *Nat. Rev. Mater.*, 2017, **2**, 1–16.
11 H. Mashhadimoslem, M. A. Abdol, P. Karimi, K. Zanganeh, A. Shafeen, A. Elkamel and M. Kamkar, *ACS Nano*, 2024, **18**, 23842–23875.
12 K. Sumida, D. L. Rogow, J. A. Mason, T. M. McDonald, E. D. Bloch, Z. R. Herm, T.-H. Bae and J. R. Long, *Chem. Rev.*, 2012, **112**, 724–781.
13 J. Van Herck, M. V. Gil, K. M. Jablonka, A. Abrudan, A. S. Anker, M. Asgari, B. Blaiszik, A. Buffo, L. Choudhury and C. Corminboeuf, *Chem. Sci.*, 2025, **16**, 670–684.
14 T. D. Burns, K. N. Pai, S. G. Subraveti, S. P. Collins, M. Krykunov, A. Rajendran and T. K. Woo, *Environ. Sci. Technol.*, 2020, **54**, 4536–4544.
15 H. Lyu, Z. Ji, S. Wuttke and O. M. Yaghi, *Chem*, 2020, **6**, 2219–2241.
16 S. Park, B. Kim, S. Choi, P. G. Boyd, B. Smit and J. Kim, *J. Chem. Inf. Model.*, 2018, **58**, 244–251.
17 Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich and M. Tsotsalas, *Angew. Chem., Int. Ed.*, 2022, **61**, e202200242.
18 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, *OpenAI blog*, 2019, **1**(8), 9.
19 M. Rahimi, S. M. Moosavi, B. Smit and T. A. Hatton, *Cell Rep. Phys. Sci.*, 2021, **2**, 100396.
20 P. Zhong, B. Deng, T. He, Z. Lun and G. Ceder, *Joule*, 2024, **8**, 1837–1854.
21 T. Guo, B. Nan, Z. Liang, Z. Guo, N. Chawla, O. Wiest and X. Zhang, *Adv. Neural Inf. Process. Syst.*, 2023, **36**, 59662–59688.
22 Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes and O. M. Yaghi, *Angew. Chem., Int. Ed.*, 2023, **62**, e202311983.
23 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.
24 J. Li, W.-J. Li, S.-C. Xu, B. Li, Y. Tang and Z.-F. Lin, *Inorg. Chem. Commun.*, 2019, **106**, 70–75.
25 J. F. Kurisingal, Y. Rachuri, Y. Gu, G.-H. Kim and D.-W. Park, *Appl. Catal., A*, 2019, **571**, 1–11.
26 A. M. Varghese, K. S. K. Reddy, N. Bhoria, S. Singh, J. Pokhrel and G. N. Karanikolos, *Chem. Eng. J.*, 2021, **420**, 129677.
27 A. Policicchio, Y. Zhao, Q. Zhong, R. G. Agostino and T. J. Bandosz, *ACS Appl. Mater. Interfaces*, 2014, **6**, 101–108.
28 B. Szczęśniak and J. Choma, *Microporous Mesoporous Mater.*, 2020, **292**, 109761.
29 F. Xu, Y. Yu, J. Yan, Q. Xia, H. Wang, J. Li and Z. Li, *Chem. Eng. J.*, 2016, **303**, 231–237.
30 W. Huang, X. Zhou, Q. Xia, J. Peng, H. Wang and Z. Li, *Ind. Eng. Chem. Res.*, 2014, **53**, 11176–11184.
31 M. Ding, R. W. Flaig, H.-L. Jiang and O. M. Yaghi, *Chem. Soc. Rev.*, 2019, **48**, 2783–2828.
32 J. H. Cavka, S. Jakobsen, U. Olsbye, N. Guillou, C. Lamberti, S. Bordiga and K. P. Lillerud, *J. Am. Chem. Soc.*, 2008, **130**, 13850–13851.
33 M. Kandiah, M. H. Nilsen, S. Usseglio, S. Jakobsen, U. Olsbye, M. Tilset, C. Larabi, E. A. Quadrelli, F. Bonino and K. P. Lillerud, *Chem. Mater.*, 2010, **22**, 6632–6640.
34 Q. Yang, A. D. Wiersum, P. L. Llewellyn, V. Guillerm, C. Serre and G. Maurin, *Chem. Commun.*, 2011, **47**, 9603–9605.
35 K. A. Adegoke, K. G. Akpomie, E. S. Okeke, C. Olisah, A. Malloum, N. W. Maxakato, J. O. Ighalo, J. Conradie, C. R. Ohoro and J. F. Amaku, *Sep. Purif. Technol.*, 2024, **331**, 125456.
36 Q. Liu, L. Ning, S. Zheng, M. Tao, Y. Shi and Y. He, *Sci. Rep.*, 2013, **3**, 2916.
37 S. Klokic, B. Marmiroli, G. Birarda, F. Lackner, P. Holzer, B. Sartori, B. Abbasgholi-Na, S. Dal Zilio, R. Kargl and K. Stana Kleinschek, *Nat. Commun.*, 2025, **16**, 7135.
38 W. L. Queen, M. R. Hudson, E. D. Bloch, J. A. Mason, M. I. Gonzalez, J. S. Lee, D. Gygi, J. D. Howe, K. Lee and T. A. Darwish, *Chem. Sci.*, 2014, **5**, 4569–4581.
39 G. Han, Q. Qian, K. Mizrahi Rodriguez and Z. P. Smith, *Ind. Eng. Chem. Res.*, 2020, **59**, 7888–7900.
40 E. Haldoupis, J. Borycz, H. Shi, K. D. Vogiatzis, P. Bai, W. L. Queen, L. Gagliardi and J. I. Siepmann, *J. Phys. Chem. C*, 2015, **119**, 16058–16071.
41 N. V. Maksimchuk, O. V. Zalomaeva, I. Y. Skobelev, K. A. Kovalenko, V. P. Fedin and O. A. Kholdeeva, *Proc. R. Soc. A*, 2012, **468**, 2017–2034.
42 H. R. Mahdipoor, R. Halladj, E. G. Babakhani, S. Amjad-Iranagh and J. S. Ahari, *RSC Adv.*, 2021, **11**, 5192–5203.
43 J.-R. Li, R. J. Kuppler and H.-C. Zhou, *Chem. Soc. Rev.*, 2009, **38**, 1477–1504.
44 C. Wang, X. Liu, N. K. Demir, J. P. Chen and K. Li, *Chem. Soc. Rev.*, 2016, **45**, 5107–5134.
45 D. DeSantis, J. A. Mason, B. D. James, C. Houchins, J. R. Long and M. Veenstra, *Energy Fuels*, 2017, **31**, 2024–2032.