Post-Pretraining Large Language Model Enabled Reverse Design of MOFs for Hydrogen Storage

Zhimeng Liu [a, 1], Yuqiao Su [a, 1], Yujie Guo [a], Jing Lin [a], Shulin Wang [a], Zuoshuai Xi [a], Zeyang Song [a], Hongyi Gao [a*], Lei Shi [b*] Ge Wang [a*]

[a] *Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Key Laboratory of Function Materials for Molecule & Structure Construction, School of Materials Science and Engineering, University of Science and Technology Beijing, Beijing 100083, PR China*

[b] *University of Science and Technology Beijing, Beijing, PR China*

[*]*Corresponding author. E-mail address: hygao@ustb.edu.cn, leishi@buaa.edu.cn, gewang@ustb.edu.cn*

Abstract. Large language models (LLMs) have achieved remarkable performance in general domains, they still face significant challenges when applied to specialized problems in fields like materials science. In this study, we enhance the performance of LLMs in the specific field of metal-organic frameworks (MOFs) for hydrogen storage by employing a post-pretraining approach to customize the LLM with domain-specific learning. By incorporating a comprehensive dataset comprising more than 2,000 MOF structures, over 7,000 related scientific papers, and a corpus exceeding 210 million tokens of specialized materials and chemical knowledge, we developed a domain-specific LLM for MOFs, referred to as MOFs-LLM. Through supervised fine-tuning, we unlocked the potential of MOFs-LLM in various tasks, including performance prediction, inverse design, mechanistic studies and application prospect analysis, with a specific focus on hydrogen storage material design challenges. In the practical application of reverse design, we utilize MOFs-LLM to mutate numerous ligands and select suitable building blocks, resulting in a structural space encompassing more than 100,000 MOFs. A MOF structure with highly promising hydrogen storage performance was ultimately successfully identified. This work effectively demonstrates the successful

application of LLMs in a specific material science domain and provides a methodological pathway that can serve as a valuable reference for future research.

Keywords: Large language model, Metal-organic frameworks, Hydrogen storage, Reverse design, Post-Pretraining.

1 Introduction

Metal-organic frameworks (MOFs) [1-3] are a recent class of hybrid porous materials composed of inorganic metal clusters or ions coordinated with organic linkers, which have gained recognition as one of the ten chemical innovations with transformative potential across various industries. The exceptional characteristics of MOFs, including their high surface areas, permanent porosity, and versatile functionalities, make them promising materials for addressing the demanding requirements of hydrogen storage, a critical component of future energy systems and clean energy technologies. The investigation into MOFs for hydrogen storage began with the synthesis of MOF-5 by Yaghi in 2004, which paved the way for exploring the potential of MOFs for hydrogen storage. Significant strides have been made in the design of high-performance MOFs for hydrogen storage, resulting in numerous achievements [4]. The diverse composition and vast number of MOFs challenge traditional trial-and-error experimentation in identifying promising MOFs that can surpass existing performance limits.

Considerable efforts have been devoted to overcoming this challenge by adopting advanced computational methods, high-throughput screening techniques [5], and machine learning algorithms [6-7]. These approaches enable the exploration of a vast chemical space of MOFs, accelerating the discovery and design of MOFs with enhanced hydrogen storage properties. However, these methods primarily rely on screening optimal materials or mining structure-property relationships from extensive data. Due to the lack of in-depth understanding, they may have limitations in perceiving or inferring certain non-quantifiable scientific principles or limitations.

Emerging large language models (LLMs) like ChatGPT [8], LaMDA [9], and ERNIE [10] have presented a new opportunity to address the abovementioned challenges. Their ability to comprehend complex information, integrate domain knowledge, and learn from human feedback holds promise for advancing the understanding and design of MOFs with enhanced hydrogen storage properties. For instance, Yaghi and colleagues [11] integrated GPT-4 into the iterative process of reticular chemistry experiments, establishing an interactive workflow between GPT-4 and human researchers. Smit and others [12] have fine-tuned language interaction for classification, regression, and molecular generation formula tasks in chemistry and materials science, surpassing the limitation of small data volume and achieving prediction accuracy comparable to dedicated machine learning methods or even better. Although LLMs have made significant progress in some applications, their knowledge and capabilities are primarily acquired during pre-training, and general-purpose models still lack the specialized expertise required for MOFs as hydrogen storage materials.

Herein, this study aims to develop a dedicated LLM tailored for the research of MOFs materials, explicitly focusing on reverse structure design for high-performance hydrogen storage materials. In particular, we have created the initial corpus that comprehensively describes the structure of MOFs materials using natural language, consisting of approximately 2,000 structural descriptions and incorporates insights from over 7,000 research papers on MOFs. We conducted post-pretraining [13] on the ERNIE [14] model and acquired over 210 million MOFs material and chemical knowledge corpus, which were utilized to develop a specialized LLM for MOFs (MOFs-LLM). Subsequently, guided by MOFs-LLM, we mutated a large number of MOFs ligands and carefully selected appropriate building blocks, thus constructing a chemical space of MOFs tailored for high-performance hydrogen storage. Finally, employing Monte Carlo calculations along with encoding and decoding strategies, we successfully reverse-engineered high-performance hydrogen storage MOFs that surpass the materials found in the CoRE-MOFs (Computation-Ready Experimental MOFs) [15] database. This study establishes

a flexible framework for the reverse design of MOFs for hydrogen storage, significantly expediting the process of MOF design and highlighting the untapped potential of LLMs in the field of materials science.

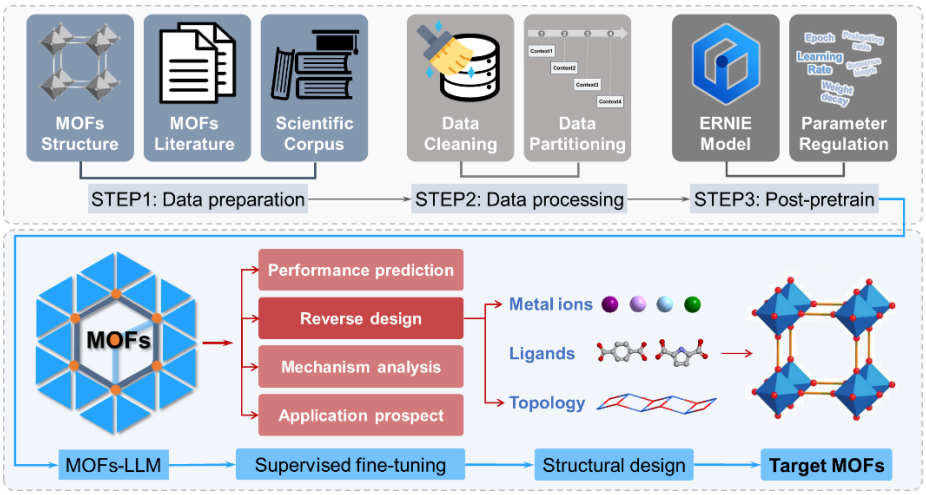

Figure 1. Flow chart of MOFs material reverse design driven by MOFs-LLM

2 Methods

2.1 Background knowledge of ERNIE Speed and Post-pretrain

ERNIE Speed [16-17] is a knowledge-enhanced language model developed by Baidu, which combines deep learning with knowledge graph technology to provide deep semantic analysis and precise text expression. Regarding architectural design, ERNIE Speed adopts an efficient transformer architecture and incorporates a knowledge enhancement mechanism. It effectively manages long-range dependencies in text through self-attention mechanisms, enabling precise understanding of complex sentence structures and generating sentences that are highly coordinated with the given context.

During the training process, ERNIE Speed utilises unsupervised learning techniques on a vast corpus of textual data for pre-training. By profoundly exploring language patterns, structures, and representations, the model can capture deep semantic information from text, enhancing its text understanding and generation capabilities. This pre-training process enables ERNIE Speed to excel in various natural language processing tasks, including text classification and question answering. Based on the pre-trained model, the post-pretrain model can be further pre-trained by Soft Prompt

Fine-Tuning (SFT) to fine-tune for different downstream tasks, thus improving the model's performance on a specific task. Furthermore, ERNIE Speed exhibits strong transfer learning capabilities. It can effectively apply knowledge learned from one task to other tasks, demonstrating excellent generalization performance. This ability enables ERNIE Speed to adapt to different domains and scenarios, providing flexible and efficient solutions for various practical applications.

Post-pretraining [18-19], an advanced transfer learning method, offers a solution for LLMs to delve into the field of materials. It involves further training a pre-trained model on a large amount of unsupervised domain data to better grasp the complex knowledge and requirements of specific tasks in that field, enhancing the model's professionalism and practicality. Through leveraging the robust logical reasoning and cross-domain synthesis capabilities inherent in LLMs, targeted post-pretraining can be utilized to integrate the intricate knowledge within the MOFs domain, potentially heralding a pattern shift in the design of MOF materials [20-21].

2.2 MOFs Material structure-hydrogen Storage properties Corpus

Through manual statistical methods, we accurately annotated the metal coordination number, ligand count surrounding the metal, and the cluster count of metals surrounding the ligands in 2000 CIF files from the CoRE-MOFs dataset. Using a directed acyclic graph script, we segmented the ligands and recorded their Simplified molecular input line entry system (SMILES) representation, molecular weight, Hydrophobic parameter calculation reference value (XLogP3), hydrogen bond donor count, hydrogen bond acceptor count, rotatable bond count, and molecular structural description from PubChem (https://pubchem.ncbi.nlm.nih.gov/). This approach helps better understand the characteristics of the ligands. We utilized Zeo++ software to compute essential indicators for understanding MOFs structure, pore characteristics, and performance, such as unit cell volume, single crystal density, gravimetric accessible surface area, accessible volume fraction, probe occupiable volume fraction, global cavity diameter, pore limiting diameter, and largest cavity diameter. Using the pymatgen[22] script, we compiled statistics on the types of metals and their coordination,

bond lengths, molecular formulas, crystal systems, and space groups in MOFs. We obtained the MOFs' topology using their MOFID [23]. By cross-referencing MOFs with Cambridge Crystallographic Data Centre (CCDC, https://www.ccdc.cam.ac.uk [24]) identifiers, we found corresponding synthesis literatures to extract framework introductions. It is important to note that our data collection is an ongoing effort, continually expanding to support future research and analysis in the field.

2.3 Fine-tune the LLM to perform specific tasks for MOFs hydrogen storage

The LLM was fine-tuned using the MOFs material structure-hydrogen storage properties corpus we prepared, enabling it to incorporate and utilize external information. [25] The LLM's comprehension ability was used to understand the input instructions related to MOFs. LLM-oriented hints were then constructed based on the instructions and the reference information to motivate the LLM to learn relevant knowledge and practical application. By employing this technique, the LLM dynamically selected reference information from the existing MOFs system and generated responses, effectively enhancing its ability to utilize external information. This process resulted in the development of the MOFs-LLM, which combines the strengths of the LLM and the specific domain knowledge of MOFs.

A multi-strategy fusion approach was implemented to address the limitations of single fine-tuning strategy in LLMs, which often suffer from low learning efficiency and poor convergence. This fusion technique combines supervised fine-tuning, contrastive learning, and reinforcement learning strategies to enhance the LLM's performance. By integrating these different fine-tuning techniques, the LLM-generated content becomes more relevant to the field of MOFs. Firstly, the supervised fine-tuning technique is employed to enable the LLM to learn specific tasks within the domain of MOFs, such as generating compliant substance structures, substance synthesis strategies, and other related tasks, utilizing labeled data as direct guidance. Secondly, contrast learning is utilized to enhance the LLM's ability to differentiate and make judgments. By exposing the model to contrasting examples and encouraging it to capture subtle differences and patterns, the LLM becomes more adept at understanding

the nuances within the MOFs domain. Finally, reinforcement learning is applied to further enhance the quality and adaptability of the LLM's generation. Through the establishment of artificial reward signals, the LLM's tendency to produce incorrect answers is reduced, while the content of correct answers is reinforced. This iterative process helps improve the overall precision and reliability of the LLM's generated outputs.

2.4 Ligand variation methods and data sources

We employed the ligand dataset established by Yaghi et al. as the cue engineering corpus. [26] This dataset encompasses 11,806 distinct molecular representations and encapsulates 3,943 distinct structural transformations. By employing supervised fine-tuning techniques, the MOFs-LLMs were refined, culminating in establishing a ligand space tailored for the reverse design of ligand variations to enhance MOFs performance.

3 Results and discussion

3.1 Data collection and processing for post-pretraining

The dataset used for post-pretraining the LLM consisted of various components. Firstly, it included a corpus of 2000 natural language descriptions specifically focused on the structure-hydrogen storage properties of MOFs. [27] Additionally, the dataset comprised over 7000 research papers related to the topic, with a subset of 800 articles (referred to as MOFs_800) highly relevant to hydrogen storage. Lastly, a generalized corpus was also incorporated as part of the training data.

MOFs structure corpus. MOFs are formed by the coordination self-assembly of metal ions with organic ligands, and their structures play a pivotal role in determining their physical and chemical properties. Key properties influenced by the MOFs structure include pore size, surface area, and pore structure. which directly affect the performance of MOFs in applications such as adsorption, energy storage, catalysis, etc. The structure of MOFs encompasses not only their chemical compositions but also their microscopic arrangements and crystal structures. It is worth noting that a LLMs may possess knowledge of the general concept of MOF composition but lacks specific

knowledge of their individual structures. Therefore, providing a clear and detailed description of MOFs' structures can greatly enhance the language model's understanding of these materials and their properties. By incorporating information about the structure of MOFs, language models can offer more accurate guidance in predicting their performance, designing novel materials, or optimizing existing ones.

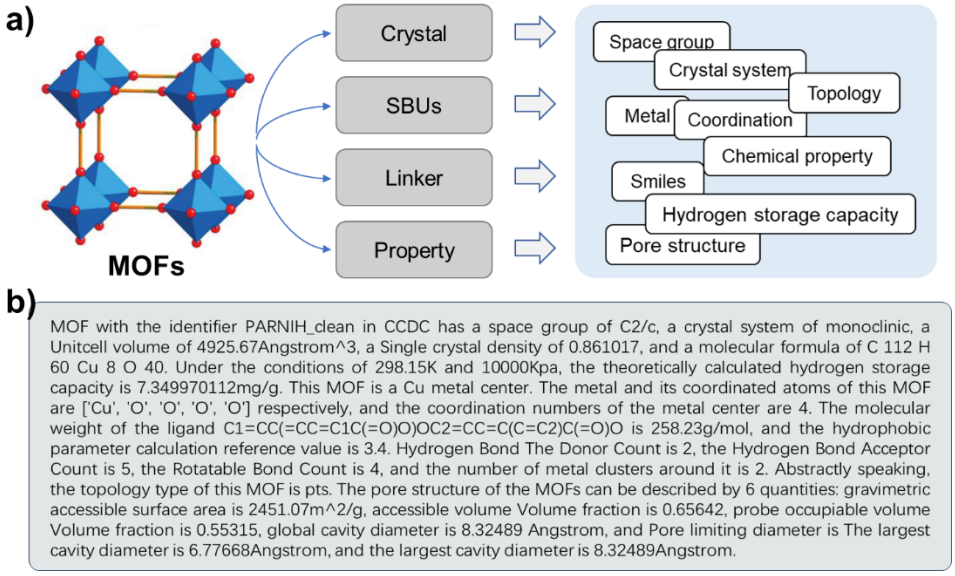

Figure 2. A natural language description of the MOFs structure, a) the elements that make up the description, and b) a sample description

As is shown in figure 2, the description of the metal in the MOF structure involved manual labeling of various attributes. These attributes include the metal type, coordination number, bond length, bond angle, and conformation of the metal center, as described in the corresponding literature. For the ligand in the MOF structure, the description encompassed several components. These components were marked based on the SMILES code, the property associated with the ligand, the molecular structural description of the ligand itself, and the coordination mode of the ligand, as mentioned in the relevant paper. Regarding the MOF framework, the description entailed additional annotations. These annotations comprised the crystal system, space group, topology, single crystal density, gravimetric surface area, volumetric surface area, void fraction, pore volume, largest cavity diameter, pore limiting diameter, and information regarding the skeleton introduction, all of which were detailed in the corresponding paper. Firstly, 11,660 MOFs in the CoRE-MOFs dataset were clustered analysis with

eight parameters, including single crystal density, gravimetric surface area, volumetric surface area, void fraction, pore volume, largest cavity diameter, pore limiting diameter, and hydrogen storage performance. Then a subset of 2,000 samples was extracted from the clustered MOFs dataset to serve as a basis for generating natural language descriptions of the MOFs' structures which aimed to capture the essential characteristics and features of the MOFs in a textual format.

Literature corpus. Using "MOF" and "hydrogen storage" as keywords, about 800 highly relevant papers in recent years were downloaded from the Web of Science. The selected papers are listed in Table S1. Another 6,563 papers were downloaded from papers covered in CoRE-MOFs. Each paper's title, abstract, chapter title and paragraph were cleaned, filtered and formatted, and 6,221 papers of higher quality were used for post-pretraining of the LLM. Segmentation of free-text paragraphs with sections from the corpus was done using a sliding window to input the model. As shown in Table 1, the literature has been processed to a level of perplexity comparable to the general scientific corpus, which proves that the quality of data cleaning is good.

Table 1. Statistics of corpus data processing results

	Number of tokens	Word repetition rate	Special character rate	Perplexity
MOFs_ALL	51580917	**0.224**	0.286	582.285
ChemSum	56009510	0.265	0.219	580.725
peS2o	**103732381**	0.304	**0.187**	**298.992**
MOFs_800	8588503	0.277	0.250	502.900

Generalized Scientific Corpus. The peS2o [28] dataset and ChemSum Datasets [29] were used as a general corpus, which had been well-cleaned of data. The ChemSum dataset is a dataset of pure chemistry by accessing the open-access chemistry scholarly journal list, a dataset with 505,548 articles on pure chemistry. The Pes20 dataset has a short corpus of about 40 million creative open-access academic articles, and we selected the material portion. The ratio of unique corpus to general corpus is 1:3, which is cleaned, filtered, and formatted for the post-pretraining of LLM to gain more

fundamental knowledge of chemistry materials and logic.

3.2 Post-pretraining and supervised fine-tuning of MOFs-LLMs

The model's knowledge base and computational power are initially shaped during the pre-training phase. However, merely fine-tuning a generic model may not fully unlock its potential within a specific domain due to limitations. Fortunately, the ERNIE model opens up the Post-pretrain API, which empowers the model with the ability to learn and gain a deeper understanding of domain-specific knowledge, including as specific terms, concepts, patterns, laws, and so on, related to MOF materials. This characteristic enables the model to become more specialized and valuable, as it gains expertise in the domain of MOFs.

Solving the LLM problem within the domain of MOFs presents significant challenges, but it is a creatively important endeavor. To address this, a post-pretraining process was conducted on the ERNIE speed model, pushing its boundaries by training it on a combined corpus of MOFs, chemistry, and materials science, totaling up to 210 million tokens. This extensive training endowed the model with a profound understanding of MOFs and the capability to tackle problems specific to this field. As a result, a large-scale prognostic model named MOFs-LLM was developed exclusively for MOFs.

This initiative would significantly improve the performance of various tasks related to MOFs, including predicting performance, reverse design, mechanism analysis and prospective studies. To ensure a reliable training process and avoid overfitting, a relatively conservative strategy was adopted. First, the 210 million corpora was employed for Post-pretrain, utilizing a relatively low learning rate. At this stage, the loss of the model slowly decreased, indicating that it was continuously migrating towards MOFs and material domains. Subsequently, an incremental training method was implemented to further train the MOFs-LLM model. This involved incorporating additional data from over 7,000 MOFs literature sources to enhance the model's performance. Research has demonstrated that continued training on datasets from the target domain can be a cost-effective approach to improve results. [30]

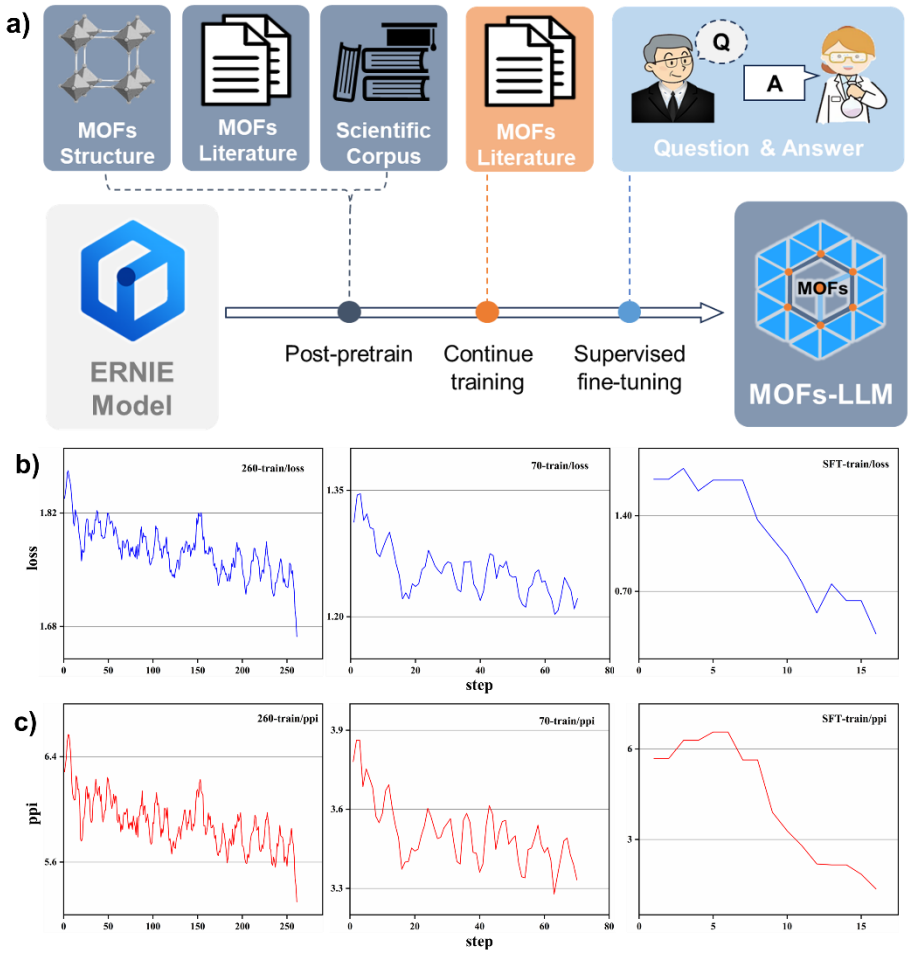

Figure 3. Post-pretrain and SFT process, a) overall training flow chart; b) changes in Loss during three training sessions; c) the changes of perplexity (ppi) during the three training sessions;

It was noticed that both the loss and the perplexity of the model exhibited significant decreases during the incremental training process (Figure 3). Although the lowest achieved loss was approximately 1.2, it would be a fairly low value for a LLMs. Perplexity, on the other hand, could be considered as a measure of a LLM's ability to predict language samples. Lower perplexity values indicate that the model performs better in the reasoning process, demonstrating higher accuracy. [31] The perplexity of MOFs-LLM dropped to approximately 3.3, indicating that the model is capable of selecting the appropriate response direction without becoming disoriented when solving problems. The reduction of loss and perplexity showed that MOFs-LLM model achieved a significant improvement in domain accuracy and effectiveness[32]. The improved performance of the MOFs-LLM model is not only reflected in its ability to

recognize a large number of samples, but more importantly, in its predictions aligning with the general rules and practical requirements within the field of MOFs.

Due to lacking of precise evaluation criteria, the efficiency evaluation of "post-pretrain" usually depends on the subsequent Supervised Fine-Tuning (SFT) process. It is possible to evaluate the performance of "post-pretrain" model on specific tasks through SFT, and improve and optimize the model training process in data science. The SFT of MOFs-LLM was executed, involving four different types of problem scenarios, including hydrogen storage performance prediction, reverse design, mechanism study, and application prospect analysis.

The advantages of the MOFs-LLM model were demonstrated in vertical experiments with ERNIE and GPT 3.5. To evaluate the performance of the models during the SFT process, a panel of three experts specializing in the field of hydrogen storage designed a set of questions. The teams led by these experts provided multiple specific and comprehensive answers. The responses generated by the three models (MOFs-LLM, ERNIE 3.5, and GPT 3.5) for 15 questions were presented to the experts, who then evaluated them based on criteria such as relevance, comprehensiveness, depth, and enlightening insights. The experts rated the answers on a scale of 1 to 5, where a higher rating indicated a more favorable perception of the responses

Table 2. Comparison of expert ratings of three LLMs

	Relevance	Comprehensiveness	Depth	Insightfulness
GPT 3.5	3.79	3.55	3.45	3.48
ERNIE 3.5	3.57	3.17	3.17	3.19
MOFs-LLM	**4.05**	**3.90**	**3.98**	**3.88**

The results, as shown in Table 2, reveal that the scores of the three models were almost equal in terms of accuracy and comprehensiveness, while the MOFs-LLM model outperformed the other two models in terms of illuminating and in-depth. It illustrated that MOFs-LLM learnt more underlying logical thinking in post-pretrain and was fully activated in the specific case of SFT. Compared to its competitors, MOFs-

LLM demonstrated a deeper level of thinking, provided more concrete answers, and supported its arguments with extensive examples. These characteristics surpassed the quality of the expert Q&A knowledge that was initially provided to the model during the SFT training process (see SI for details). Despite using a relatively small amount of SFT data, the data that was utilized was of high quality. This allowed the model to be finely tuned and achieve excellent performance, potentially even leading to innovative conclusions. It is possible that even with as little as 0.5% of the available data, the model could still exhibit impressive results. In the following reverse design process, MOFs-LLM was employed to mutate the ligands and assisted with the selection of suitable metals and topologies as building blocks, which aimed to explore the relative relationship between MOFs materials and their hydrogen storage properties.

3.3 Reverse design of high-performance hydrogen storage MOFs material

One of the keys in reverse design is the construction of a vast chemical space to cover a wide range of possible MOFs material structures and properties. In this way, a systematic approach enables the rapid screening and design of MOFs materials with targeted properties according to the specific requirements. While the types of metals and topologies are largely defined, the ligands are the primary targets for modulation with unlimited design potential. The structure and functional groups of the ligands allow the control of critical features such as pore size, surface area, and pore volume of MOFs materials, which in turn affects their hydrogen storage properties. Therefore, the precise regulation and optimization of the hydrogen storage properties of MOFs materials can be achieved through the rational design and variation of ligands.

To construct the extensive chemical space required for reverse design of MOFs materials, the ligands were subjected to mutation using the MOFs-LLM model (see Fig. 4). The approach of SFT via APIs was employed, which is particularly suitable for functional implementations with clearly defined requirements, as opposed to pre-training on large-scale unsupervised text data. It is expected that MOFs-LLM would perform as many mutations as possible on the given ligands under consideration of symmetry to explore its ability to analyze the chemical space of MOFs. Simultaneously,

the SMILES data with Pubchem CID were used for the next round of mutations each time to improve the synthesis ability of the mutated molecules.

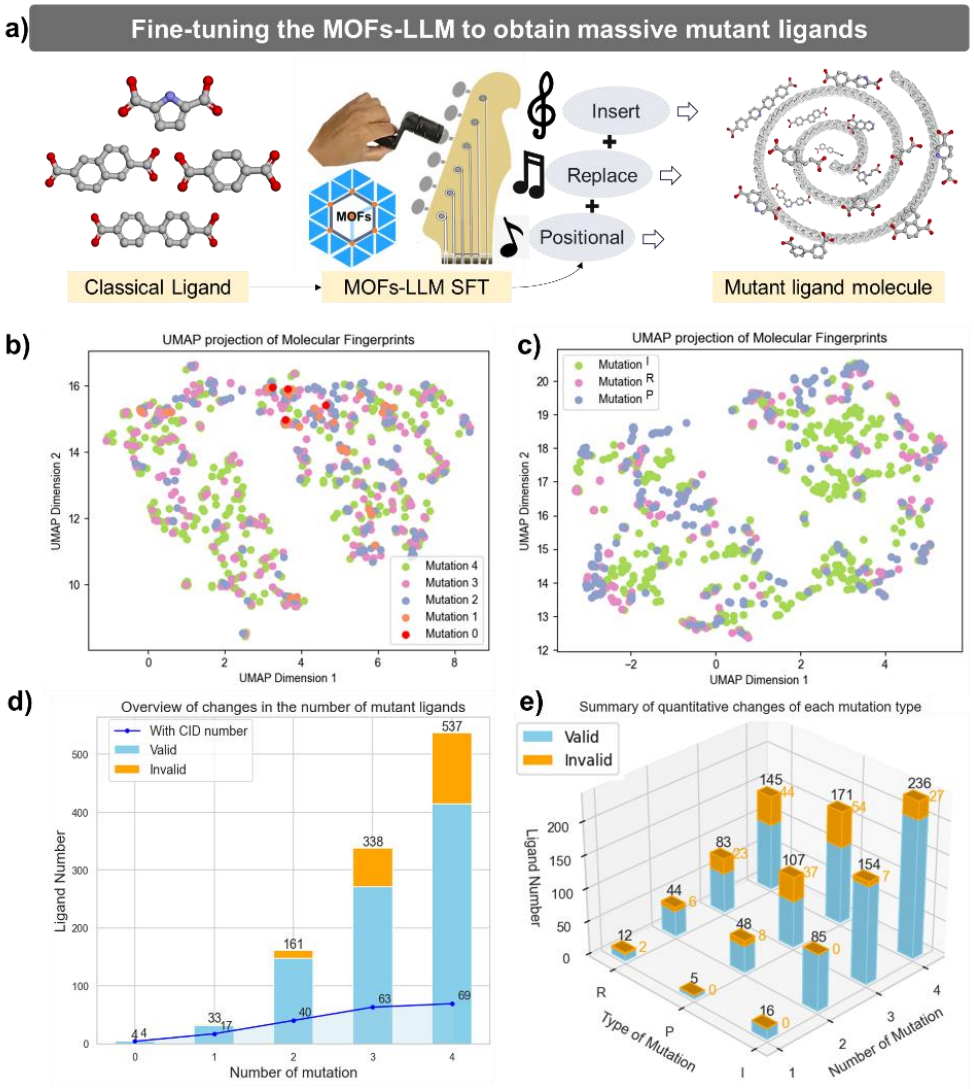

Figure 4. Fine-tune MOFs-LLM for a large number of high-quality ligands. a) flow chart of ligand variation, b) and c) distribution of the ligand was generated for each round/operation, d) and e) rationality analysis of ligands generated by each round/operation.

Four rounds of ligand mutation were performed using natural language representations (SMILES code), starting from the four most basic ligands, as shown in Fig. x. The existence of variant ligands was validated using RDKit, and their molecular fingerprints were calculated and subsequently visualized in Uniform Manifold Approximation and Projection (UMAP). [33-34] As shown in figure 4b, 4c, the number of mutant ligands steadily increased with increasing operation rounds. After only four rounds of iteration, thousands of variant products were generated from the initial four

base ligands. Among these variants, mutations performed using the insertion (I) method resulted in the highest number of mutation products, followed by the P method, and finally the R method. The insertion (I) method exhibited high versatility, allowing for the transformation of nearly any ligand into another possible ligand by inserting a benzene ring, vinyl group, or azo group. It is worth noting that as the number of rounds increased, the proportion of non-existent mutation products also increased. This phenomenon occurred because the generated structures gradually approached the limits of the model's comprehension ability.

The UMAP visualization of the distribution allowed intuitive observation of the mutation and design process of the ligands of MOFs. In UMAP, each point represented a mutation product, and the position indicated their similarity in the chemical space. By analyzing the distribution, it was evident that numerous variant products were successfully generated through several rounds of manipulation, starting from a limited number of initial essential ligands. These variant ligands exhibited distinct chemical structures and functional groups, demonstrating a rich diversity in the generated ligand space. Moreover, the UMAP points were uniformly distributed throughout the chemical space without apparent aggregation. It demonstrated that this design strategy could efficiently explore and cover all possibilities in the chemical space beyond limiting to a specific region or type. Detailed molecular maps of the ligands are available in the Supporting Information section.

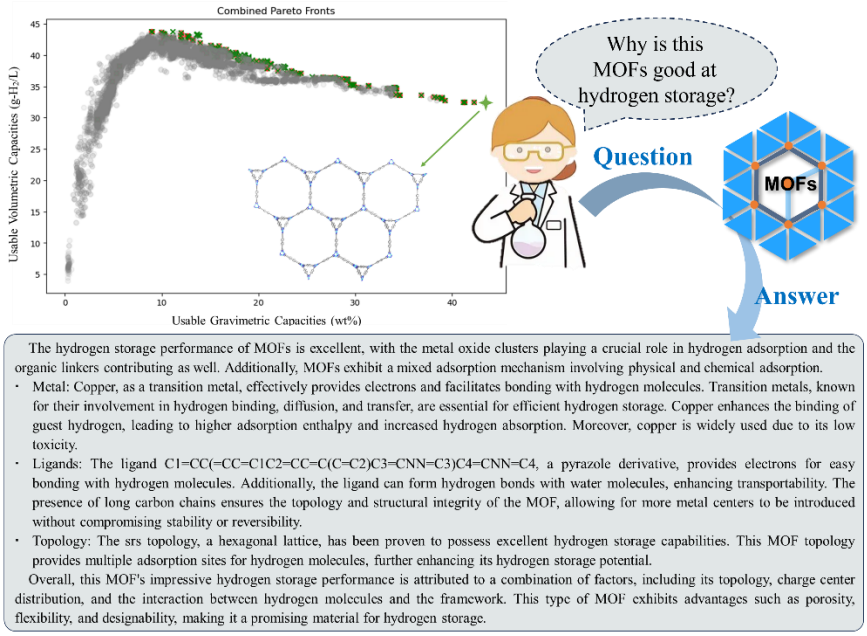

Figure 5. The structure distribution of high-performance MOFs, where the green star in the upper right is the MOFs with the highest performance in the figure, and the bottom is the explanation given by MOFs-LLM.

Through bootstrapping and questioning, MOFs-LLM selected 150 classes from the above-compiled ligands. At the same time, in parallel, 20 metals and 20 topologies were screened out as potential high-performance hydrogen storage material building blocks by MOFs-LLM, which composed 100,842 MOFs data by ToBaCCo. [35] The current pool of nodes and topologies represented a reasonable and achievable high-performance structure for the MOFs mesh framework. The coder-decoder SmVAE allows mapping the frame of MOFs with discrete representations into continuous vectors and back again. [36] The usable volumetric (UV) capacities and usable gravimetric (UG) capacities values of 3182 MOF structures consisting of the first two rounds of mutant ligands at between 77 K/100 bar (filled state) and 160 K/5 bar were predicted using the trained the highly randomized trees (ERT) as labelling data in SmVAE space. [37] As shown in Fig. X, the properties and distribution of MOFs structures in the overall space were predicted with the assistance of SmVAE, forming a chemical space of high-performance hydrogen storage MOFs beyond the CoRE-MOFs database. Since the SmVAE space was a vector space, continuous optimization and

search algorithms were used to find local minima or maxima. It is possible to sample and reconstruct the MOFs frame by decoding. A subset of the generated MOF structures exhibited hydrogen storage performance that exceeded the performance of the initial structures in the labelled data, as shown by the green crosses marking the Pareto front surface of the balanced UV and UG performances in Fig. 5. The MOF corresponding to the Pareto optimum point for balancing UG and UV was decoded, as marked by the green star in the figure 5. It is an excellent performance hydrogen storage material at between 77 K/100 bar (filled state) and 160 K/5 bar with ug of 43.433 wt% and UV of 32.447 g-H_2/L. Subsequently, the relationship between its structure and the hydrogen storage mechanism was analyzed using MOFs-LLM, aiming to elucidate the reasons behind its superior hydrogen storage capacity. The responses provided by MOFs-LLM were compiled and analyzed. MOFs-LLM responses were as follows:

4 Conclusion

To meet the specific requirements of LLMs in the field of MOFs materials, large amounts of structural and performance information on MOFs were analyzed, and 7000 pieces of scientific and technological literature in related fields were internalized, creating the first corpus describing MOFs materials in natural language. A large language model (MOFs-LLM) tailored specifically for the field of MOFs was developed using this corpus to post-pretrain ERNIE. Supervised fine-tuning of the expert Q&A on reverse design of MOFs targeting hydrogen storage materials, unleashed the model's potential in tasks involving performance prediction, reverse design, mechanistic studies, and analysis of application prospects. In the practical application of reverse design, a structured space containing more than 100,000 MOFs was constructed using MOFs-LLM to perform many mutations of ligands and selecting appropriate building blocks in combination with MOFs-LLM. Ultimately, a target MOF structure with high hydrogen storage performance was successfully identified. This work demonstrates the practical application of LLMs in specific areas of materials science and provides a methodological approach for reference.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFB3500700), National Natural Science Foundation of China (No. 52373261), Beijing Natural Science Foundation (L233011), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022A1515010185).

Conflict of Interest The authors declare no conflict of interest.

References

[1] H. Li, M. Eddaoudi, O. M. Yaghi, et al. Design and synthesis of an exceptionally stable and highly porous metal-organic framework[J]. Nature,1999, 402, 276–279.

[2] M. O'Keeffe, M.A. Peskov, O.M. Yaghi, et al. The Reticular Chemistry Structure Resource (RCSR) database of, and symbols for, crystal nets[J]. Accounts of chemical research, 2008, 41, 1782–1789.

[3] Z. Wu, Y. Li, G. Wang, et al. Recent advances in metal-organic-framework-based catalysts for thermocatalytic selective oxidation of organic substances[J]. Chem Catalysis, 2022, 2, 1009–1045.

[4] S.M. Moosavi, A. Nandy, B. Smit, et al. Understanding the diversity of the metal-organic framework ecosystem[J]. Nature Communications, 2020, 11, 4068.

[5] K.T. Butler, D.W. Davies, A. Walsh, et al. Machine learning for molecular and materials science[J]. Nature, 2018, 559, 547–555.

[6] K.M. Jablonka, D. Ongari, B. Smit, et al. Big-Data science in porous materials: Materials genomics and machine learning[J]. Chemical Review, 2020, 120, 8066–8129.

[7] S.M. Moosavi, K.M. Jablonka, B. Smit, et al. The role of machine learning in the understanding and design of materials[J]. Journal of the American Chemical Society, 2020, 142, 20273–20287.

[8] OpenAI, J. Achiam, S. Jain, et al. GPT-4 Technical Report[J]. Computation and Language, 2023.

[9] R. Thoppilan, D. D. Freitas, Q. Le, et al. LaMDA: Language Models for Dialog Applications[J]. Computation and Language, 2022.

[10] Y. Sun, S. Wang, H. Wu, et al. Ernie: Enhanced representation through knowledge integration[J]. Computation and Language, 2019.

[11] Z. Zheng, Z. Rong, O. M. Yaghi, et al. A GPT-4 Reticular Chemist for Guiding MOF Discovery[J]. Angewandte Chemie International Edition, 2023, 62, e202311983.

[12] K. M. Jablonka, P. Schwaller, Berend Smit, et al. Leveraging large language models for predictive chemistry[J]. Nature Machine Intelligence, 2024, 6, 161–169.

[13] L. Pan, C. Hang, M. Yu, et al. Multilingual BERT Post-Pretraining Alignment[J]. Association for Computational Linguistics, 2021, 210–219.

[14] X. Fang, F. Wang, L. Song, et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model[J]. Nature Machine Intelligence, 2023, 5, 1087–1096.

[15] Y. G. Chung, E. Haldoupis, R. Q. Snurr, et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal–Organic Framework Database: CoRE MOF 2019[J]. Journal of Chemical & Engineering Data, 2019, 64, 5985–5998

[16] Y. Sun, S. Wang, Z. Liu, et al. ERNIE: Enhanced Representation through kNowledge IntEgration[J]. Computation and Language, 2020.

[17] Y. Sun, S. Wang, H, Wang, et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation[J]. Computation and Language, 2021.

[18] F. Faisal, A. Anastasopoulos, Investigating Post-pretraining Representation Alignment for Cross-Lingual Question Answering[J]. Computation and Language, 2021.

[19] X. Qiu, T. Sun, Z. Liu, et al. Pre-trained Models for Natural Language Processing: A Survey[J]. Science China Technological Sciences, 2020, 63(10), 1872-1897.

[20] J. Devlin, M. W. Chang, K. Toutanova, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [J] In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, 1, 4171-4186.

[21] A. Vaswani, N. Shazeer, I. Polosukhin, et al. Attention is All You Need[J]. In Advances in Neural Information Processing Systems, 2017, 30, 5998-6008.

[22] S. P. Ong, W. D. Richards, G. Ceder. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis[J]. Computational Materials Science, 2013,

68, 314-319.

[23] B. J. Bucior, A. S. Rosen, R. Q. Snurr, et al. Identification Schemes for Metal–Organic Frameworks to Enable Rapid Search and Cheminformatics Analysis[J]. Crystal Growth and Design, 2019, 19(11), 6682–6697.

[24] F.H. Allen, Structural Science the Cambridge Structural Database: a quarter of a million crystal structures and rising, Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials, 2002, 58, 380-388.

[25] Z. Yang, Q. Sun, S. Yuan. Advances in machine learning methods for computational design of metal–organic frameworks[J]. Coordination Chemistry Reviews, 2021, 436, 213836.

[26] Z. Zheng, A. H. Alawadhi, S. Chheda, et al. Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned GPT models[J]. Journal of the American Chemical Society, 2023, 145(51): 28284-28295.

[27] M. Zhang, Y. Huang, Z. Lin, et al. Machine learning constructs color features to accelerate development of long-term continuous water quality monitoring[J]. Journal of Hazardous Materials, 2024, 461, 0304-3894.

[28] K. Lo, L. L. Wang, D. Weld, et al. S2ORC: The Semantic Scholar Open Research Corpus[J]. Association for Computational Linguistics, 2020, 4969–4983.

[29] Gr. Adams, B. H Nguyen, N. Elhadad, et al. What are the Desired Characteristics of Calibration Sets? Identifying Correlates on Long Form Scientific Summarization[J]. Computation and Language, 2023.

[30] C. Zhou, P. Liu, O. Levy, et al. LIMA: Less Is More for Alignment[J]. Machine Learning, 2023.

[31] J. Kaplan, S. McCandlish, D. Amodei, et al. Scaling Laws for Neural Language Models[J]. Machine Learning, 2020.

[32] S. Gururangan, A. Marasovic, N. A. Smith, et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks[J]. Association for Computational Linguistics, 2020.

[33] L. McInnes, J. Healy, J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction[J]. Machine Learning, 2020.

[34] RDKit: Open-source cheminformatics. https://www.rdkit.org

[35] R. Andersona, D. A. Gómez-Gualdrón, Increasing topological diversity during

computational “synthesis” of porous crystals: how and why[J]. CrystEngComm, 2019,21, 1653-1665.

[36] A. Ahmed, D. J. Siegel, Predicting hydrogen storage in MOFs via machine learning[J]. Patterns, 2021, 2(7), 100305.

[37] Z. Yao, R. Q. Snurr A. Aspuru-Guzik, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models[J]. Nature Machine Intelligence volume 2021, 3, 76–86.