npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-025-01540-6

Construction of a knowledge graph for framework material enabled by large language models and its application

Xuefeng Bai[1], Song He[2], Yi Li[1], Yabo Xie[1], Xin Zhang[1], Wenli Du[2] ✉ & Jian-Rong Li[1] ✉

Framework materials (FMs) have been extensively investigated with a plethora of literature documenting their unique properties and potential applications. Despite this, a comprehensive knowledge graph for this emerging field has not yet been constructed. In this study, by utilizing the natural language processing capabilities of large language models (LLMs), we have established a comprehensive knowledge graph (KG-FM). It covers synthesis, properties, applications, and other aspects of FMs including metal-organic frameworks (MOFs), covalent-organic frameworks (COFs), and hydrogen-bonded organic frameworks (HOFs). The knowledge graph was constructed through the analysis of over 100,000 articles, resulting in 2.53 million nodes and 4.01 million relationships. Subsequently, its application has been explored for enhancing data retrieval, mining, and the development of sophisticated question-answering systems. Especially when integrating the KGs with LLMs, resulted Qwen2-KG not only achieves a higher accuracy rate of 91.67% in question-answering than existing models but also provides precise information sources.

Metal-organic frameworks (MOFs), first reported in 1995[1], have garnered immense interest due to their unique crystalline and porous structure formed by metal nodes and organic linkers. The discovery of MOFs marks a significant breakthrough in the field of materials science[2]. Subsequently, covalent organic frameworks (COFs)[3] and hydrogen-bonded organic frameworks (HOFs)[4] have appeared and attracted intense attention of researchers. MOFs, COFs, and HOFs were called framework materials (FMs), which exhibit greater flexibility and innovation potential in terms of structural diversity, controllable porosity, and functional modification[5–12]. These properties endow them with a wide range of potential applications in various fields including luminescence[13], sensing[14], gas storage[15], separation[16], catalysis[17], proton conduction[18], drug delivery[19], and so on[20–22]. Therefore, the design, synthesis, and application of these FMs have become a focal point for research and exploration among chemists, materials scientists, and engineers.

The knowledge graph (KG) functions as a structured repository for semantic information, using a graphical model to depict entities, concepts, and their relationships[23,24]. It has great potential application for science education and research. For example, its logical structure and visual representation clearly reveal the relationships between information, facilitating the understanding of complex concepts during students' learning[25]. Leveraging the powerful logical relationships expressed by knowledge graphs and technologies such as graph embeddings[26], knowledge graphs have played a crucial role in supporting scientific research in areas such as chemical safety[27], drug discovery[28], property prediction[26], and the expansion of chemical reactions[29]. Additionally, KGs reveal intricate relationships, aiding researchers in exploring interdisciplinary links and pinpointing potential research directions and innovation opportunities, thereby expediting scientific discovery[23,30]. Recently, the integration of KGs with large language models (LLMs) has been explored to promote the accuracy of information provided by LLMs, thereby broadening the application scope of KG in AI systems[31,32]. Despite their broad potential application, FMs have not yet benefited from a comprehensive knowledge graph[33,34].

Crafting KGs involves transforming domain expertize and comprehensive research data into interconnected knowledge networks[35]. This process typically demands significant human effort, since researchers must meticulously sort through a vast array of fragmented and unstructured information within the literature[23,24]. Furthermore, it challenges their ability to discern the logical relationships between pieces of information and to organize them into a structured and interconnected knowledge system[36]. Application automation and intelligent technologies have reduced the workload and boosted the efficiency of knowledge graph construction[37]. LLMs are capable of automatically extracting, semantically analyzing, and logically interpreting literature content. This approach has already proven

[1]Beijing Key Laboratory for Green Catalysis and Separation and Department of Chemical Engineering, College of Materials Science & Engineering, Beijing University of Technology, Beijing, PR China. [2]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai, PR China. ✉e-mail: wldu@ecust.edu.cn; jrli@bjut.edu.cn

Fig. 1 | Flowchart for constructing a knowledge graph from the literature and using it for knowledge querying and enhanced LLMs retrieval and application of knowledge graph.

its value in chemistry[30], materials[38], engineering[39], and biology[36,40], significantly reducing manual efforts in organizing and integrating information. It is expected that knowledge graph of various research fields will be soon developed with the assistance of these AI technologies.

Herein, as shown in Fig. 1, we have compiled the existing literature on MOFs, COFs, and HOFs. By employing LLMs, we organized and analyzed more than 100,000 documents in a high-throughput manner. A KG is built on the Neo4j platform including 2.53 million nodes and 4.01 million relationships. The graph focuses on FMs, encompassing their fundamental information, properties, applications, and sources. The KG can assist researchers in information retrieval, visualization, and data analysis. Additionally, by coupling with LLMs, the KG can significantly enhance the question-answering capability of LLMs in the field of framework materials. The accuracy rate of the system is 91.67%, significantly higher than that of existing LLMs (GPT-4: 33.33%). Particularly in the chain-of-thought (CoT) reasoning tasks of LLMs, the KG provides rich background information to facilitate reasoning and question-answering.

Methods

Collecting information of relevant literature

The relevant journal articles on MOFs, COFs, and HOFs published before May 8, 2024, were retrieved from the Web of Science database using the following search queries (Eqs. 1–3):

$$\mathrm{TS} = (\mathrm{MOF\ OR\ MOFs\ OR\ ``Metal\ Organic\ Framework"\ OR\ ``Metal-Organic\ Framework"}) \quad (1)$$

$$\mathrm{TS} = (\mathrm{COF\ OR\ COFs\ OR\ ``Covalent\ Organic\ Framework"\ OR\ ``Covalent-Organic\ Framework"}) \quad (2)$$

$$\mathrm{TS} = (\mathrm{HOF\ OR\ HOFs\ OR\ ``Hydrogen-bonded\ Organic\ Framework"}) \quad (3)$$

The abstracts and publication details (including DOI, authors, publication date, and journal information as shown in Table S1) of the retrieved papers were exported from Web of Science. The exported information is saved in TXT files as text.

Abstract information extraction

Information extraction from the abstracts was carried out by converting the text into a JSON format with logical relations using Qwen2-72B. The task and output format for Qwen2-72B were defined in the prompt (Figs. S1–S3). We focused on whether the nodes extracted from the abstracts and the relationships between them were accurate and comprehensive. The evaluation of the model's conversion process is similar to the evaluation of classification algorithms in machine learning (Fig. S4). Only accurate and comprehensive relationships are considered correct (True Positive, TP).

Knowledge graph construction and usage

The knowledge graph was constructed using the Neo4j software. Through the Python interface, the publication information and the JSON files generated by the LLM were imported into Neo4j. The code can be accessed at https://github.com/MontageBai/KGFM. The constructed knowledge graph was utilized by invoking the graph database via the Neo4j Docker and conducting data visualization and analysis. The version of Neo4j we used in this study is 5.12.0.

Integration of LLMs with the knowledge graph

The Retrieval-Augmented Generation (RAG) process using LLMs was divided into 3 steps, and the entire process was implemented using Python.

1. *Generate Cypher Query*: Construct a Cypher query based on the user's question to retrieve relevant information from a Neo4j database.

2. *Execute Query and Retrieve Data*: Run the generated Cypher query against the Neo4j database to obtain data related to the user's query from the KG.
3. *Formulate Answer*: Use the retrieved KG data to formulate a precise and professional answer to the user's question.

The code can be accessed at https://github.com/MontageBai/KGFM.

Versions and usage of large language models

The open-source models were accessed locally, whereas the closed-source models were accessed through web interfaces. The specific version numbers are provided in Table 1.

Table 1 | The large language model version and web site used in this study

Model	Version	Accessed URL
Qwen	Qwen2-72B-Instruct	https://huggingface.co/Qwen/Qwen2-72B-Instruct
Llama	Meta-Llama-3-70B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
GLM	glm-4-9b-chat	https://huggingface.co/THUDM/glm-4-9b-chat
Copilot	Accessed on June 6	https://copilot.microsoft.com/
GPT-4	Accessed on June 6	https://chatgpt.com/

Table 2 | The literature on building KGs and the composition of KGs

	Literature	Nodes	Relationships	Relationship Types
MOF	68400	1351477	2439563	54535
COF	21776	553888	809375	27377
HOF	15203	628341	757736	18270

Evaluation of LLMs combined with knowledge graphs

To evaluate the performance of large language models (LLMs) augmented with knowledge graphs, we designed a test set consisting of 150 questions related to framework materials. Both the questions and their reference answers can be accessed at https://github.com/MontageBai/KGFM. The evaluation of the answers was based on manual judgment; an LLM's answer was considered correct (True) only if it matched the standard answer; otherwise, it was marked as incorrect (False).

Results

Preparation of knowledge graph

We have designed and implemented a framework for the efficient construction of KGs from literature, as illustrated in Fig. 1. The framework's process comprises data preprocessing, entity and relationship recognition using LLMs, knowledge graph construction, and the application of the knowledge graph. In the data preprocessing phase, we retrieve and collected more than 100,000 published academic literature. As shown in Figs. S5–S7, the data is stored in multiple txt files in the form of text, which is convenient for subsequent retrieval by LLMs. Extracted unstructured data, particularly abstracts, require organization, extraction, and summarization to be converted into a relational database. Here, we use the Qwen2 LLM to conduct a detailed analysis of the summary, identifying key information and organizing it into nodes and relationships, which are then saved in a JSON file. As shown in Figs. S8–S10, LLM effectively recognizes critical elements (such as research methods, experimental results, theoretical concepts, etc.) in the digest and converts them into a structured JSON format using customized prompts (Figs. S1–S3). This step is critical in knowledge graph construction as it determines the quality and accuracy of the nodes and edges in the graph. Therefore, we manually reviewed 100 results extracted by LLM. Based on the concept of a confusion matrix in machine learning (Fig. S4), the model achieved a TP rate of 98% for accurate and comprehensive information extraction, with a FN rate of 2% for inaccurate and comprehensive information extraction. The F1 score is 0.9898. Overall, with the given prompt, LLM performs well in extracting information from abstracts, and this prompt was subsequently used for high-throughput extraction.

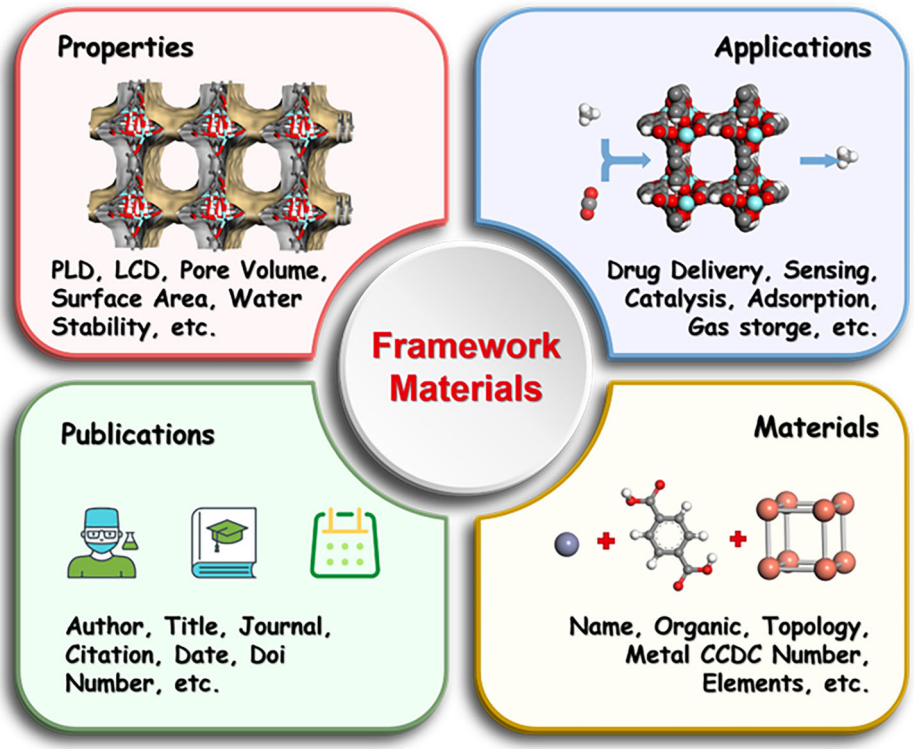

Fig. 2 | The main types of information in the knowledge graph include the application, properties, intrinsic information of the material, and related information from journal publications.

Construction of knowledge graph

In the knowledge graph construction phase, we use Cypher statements to import processed node and relationship data from the LLM, as well as structured metadata such as titles and DOI numbers, into the Neo4j graph database. Here, the generated JSON files are first converted into Cypher queries to import the data into the knowledge graph. The structured data is then linked to the nodes parsed by Qwen through manually defined relationships, such as "Derived from", "Published in (Journal)", "Published at (Date)", and etc. More details can be found in the source code at https://github.com/MontageBai/KGFM.

The knowledge graph, built from over 100,000 academic papers, encompasses over 2.53 million entities (nodes) and 4.01 million relationships, as detailed in Table 2. Among the materials, MOFs, as one of the earliest developed, hold a significant place in the knowledge graph. Figure 2 illustrates that the knowledge graph is centered on materials and includes information about their properties, structures, applications, performance, and related reports. The knowledge graph not only provides researchers with a comprehensive and multidimensional database of material information but also offers robust data support for the research and development of materials science. In the knowledge graph, each material is represented as a node, connected across multiple dimensions, including properties, synthesis methods, and application fields, forming a vast networked structure. This structured data representation enables researchers to intuitively grasp the relationships between materials and their potential value in various application scenarios.

Applications of knowledge graph

In this section, we report the exploration of KG in information retrieval as well as their integration with large language models (LLMs). The combination of KG and LLMs not only facilitates the querying KG with natural language but also addresses key challenges of LLM such as factual inaccuracies and limited domain specificity. Particularly, incorporating KGs into the chain-of-thought (CoT) reasoning of LLMs significantly enhances reasoning quality and interpretability, thereby enabling precise guidance for framework material screening.

$$\begin{aligned} &\text{MATCH(a)} - [\text{r} : \text{DRIVER\_FROM}] \\ &\quad \rightarrow (\text{b} : \text{Node name: ``Trace removal of benzene . . . ''}) \\ &\quad \text{RETURN a, r, b;} \end{aligned} \tag{4}$$

The Neo4j platform offers a user-friendly graphical interface that allows domain experts to query the knowledge graph using simple search terms. As shown in Fig. 3, we queried the work of Li and colleagues[41] using a simple Cypher statement (Eq. 4). From the knowledge graph, it is evident that the paper focuses on the material BUT-55 and reports on the application of a series of isomorphic MOFs for the trace adsorption of benzene. The structure of BUT-55 is a bi-armed structure, consisting of the metal Co and the ligand H_2BDP (1,4-di(1H-pyrazol-4-yl)benzene). The paper also explains the adsorption mechanism further by testing single crystals loaded with benzene. Similar to Fig. 3, information from different literature sources can also be obtained through Eq. S1 and Eq. S2 (Figs. S11–S12). We can also

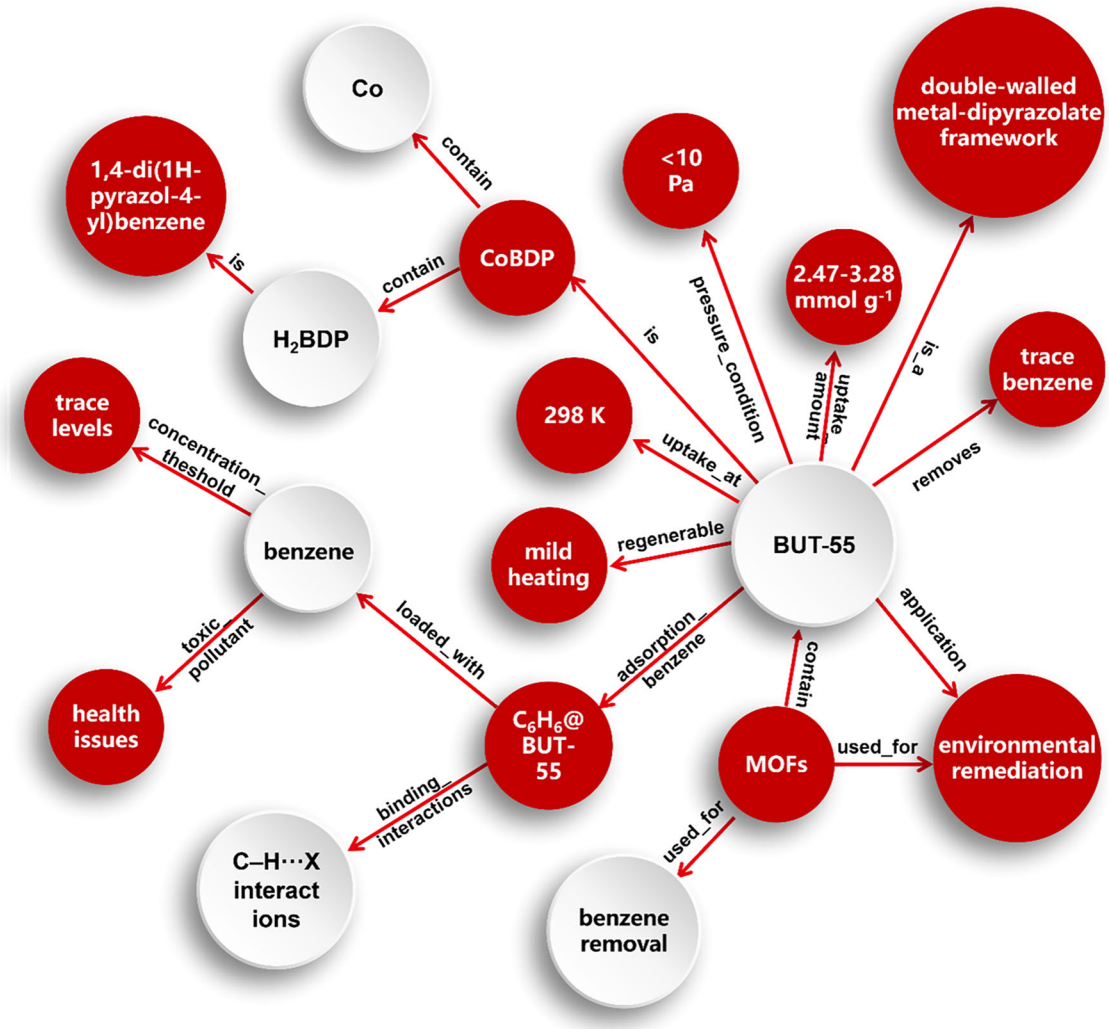

Fig. 3 | The result by using a knowledge graph to query the literature titled "Trace removal of benzene vapor using double-walled metal–dipyrazolate frameworks".

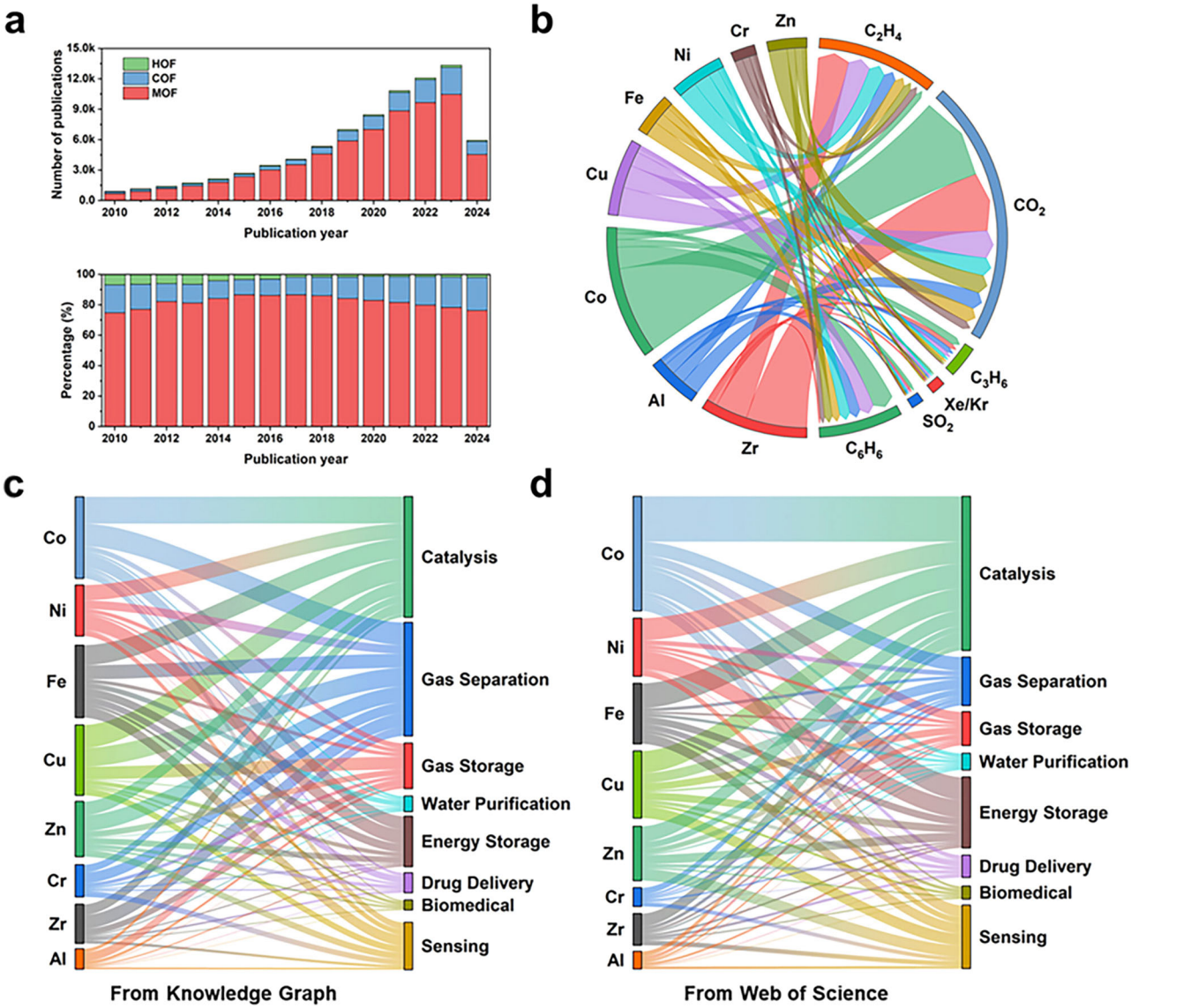

Fig. 4 | Analyzing knowledge using KGs. a Publication of literature on MOF, COF, and HOF; **b** Report of MOFs of different metals for gas adsorption; **c** Application of MOFs of different metals through knowledge mapping; **d** Application of MOFs of different metals through web of science.

conduct a material-specific search. Overall, with the help of the knowledge graph, researchers can quickly search for and access research literature on MOFs, greatly facilitating knowledge discovery and research progress in the field of materials science.

The application of KGs not only enables us to delve into potential information but also offers a clearer and more comprehensive perspective for analyzing and tracking domain trends and research focuses, surpassing traditional review papers. As shown in Fig. 4a, over the past decade, there has been a significant rise in the number of publications on MOFs, COFs, and HOFs, with a particularly notable increase in the proportion of COFs in the literature. This highlights the increasing importance and popularity of COFs in research. In Fig. 4b, we focus on the research into MOFs with different metal centers for adsorption applications. The analysis reveals that cobalt Co-based MOFs occupy a significant proportion in CO_2 adsorption studies, providing important guidance for future material design and suggesting the potential of Co-based MOFs in the field of carbon capture and conversion. Moreover, Fig. 4c and d showed the results of data mining using the knowledge graph with those from the Web of Science. Through statistical analysis using search expressions (as detailed in Table S2), we discovered significant discrepancies between the two sources. This discrepancy is likely due to the superior capability of KG to discover hidden links between different terms. For instance, in some literature, classic Cu-based MOFs such as HKUST-1 may not explicitly state the metal type in the title and abstract. The knowledge graph, however, can directly link Cu to HKUST-1, thereby enhancing the accuracy and reliability of the data mining process.

The integration of KGs with LLMs to enhance their response capabilities has been implemented across various domains[32,39]. Here, we employ the Qwen2 model as the foundation and enhance its retrieval capabilities by incorporating the FMs knowledge graph (Fig. S13). As shown in Fig. 5, we compare the Qwen2 model enhanced with the knowledge graph to other general-purpose models using the structure of BUT-55 as an example. Even with internet-accessible retrieval, the general-purpose models are unable to accurately answer queries regarding to BUT-55. This is because the original text's abstract does not mention the term BUT-55 specifically, but rather refers to it in the range of BUT-53 to BUT-58[37], which poses a challenge for general LLMs. The Qwen2 model, enhanced with the knowledge graph, accurately answers questions about the structure and properties of BUT-55 through Cypher queries. As KGs furnish publication details for literature, LLMs can now provide precise sources of information, effectively addressing the limitations of LLMs in citation and source tracing. In addition to

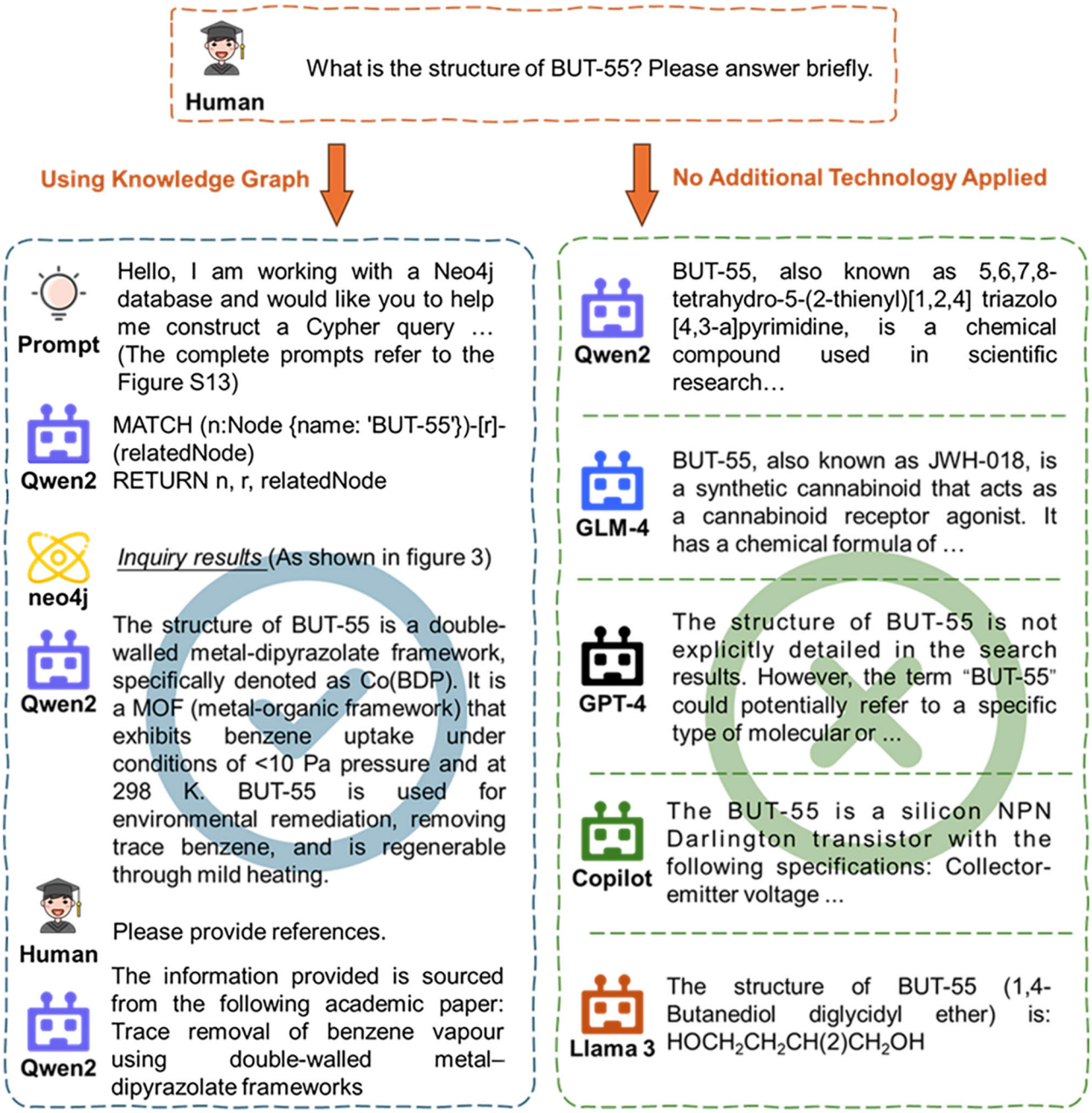

Fig. 5 | A comparison of the use of a knowledge graph to enhance the LLM question answering and the use of no techniques, taking the structure of BUT-55 as an example.

querying material structures, the integration of LLMs with the KG can also provide accurate answers of material properties (Figs. S14–S15), the content of literature (Figs. S16–S17), and material applications (Figs. S18–S19). Furthermore, it can supply specific references, either one or multiple, as needed.

Subsequently, we adopted an evaluation approach that has been previously used to assess the performance of large language models in the field of MOFs[42]. We prepared a test set consisting of 150 questions (Table S3) and tested the response accuracy of different general-purpose models as well as open-source models combined with knowledge graphs against this set (Table 3). The evaluation criteria for the accuracy of the responses were determined by two domain experts, who judged the correctness and relevance of the answers. An answer was considered correct if it provided the accurate information without any misleading content. The Qwen2-KG

Table 3 | Test general LLMs and Qwen2 enhanced with knowledge graph using 150 questions as a test set

LLMs	Accuracy	Standard error
Qwen2-KG	91.67%	0.47%
GLM4-KG	80.67%	0.94%
Llama3-KG	88.33%	1.41%
Qwen2	14.33%	0.47%
GLM4	5.67%	0.47%
Llama3	8.00%	0.94%
GPT4	33.33%	1.89%
Copilot	31.33%	2.83%

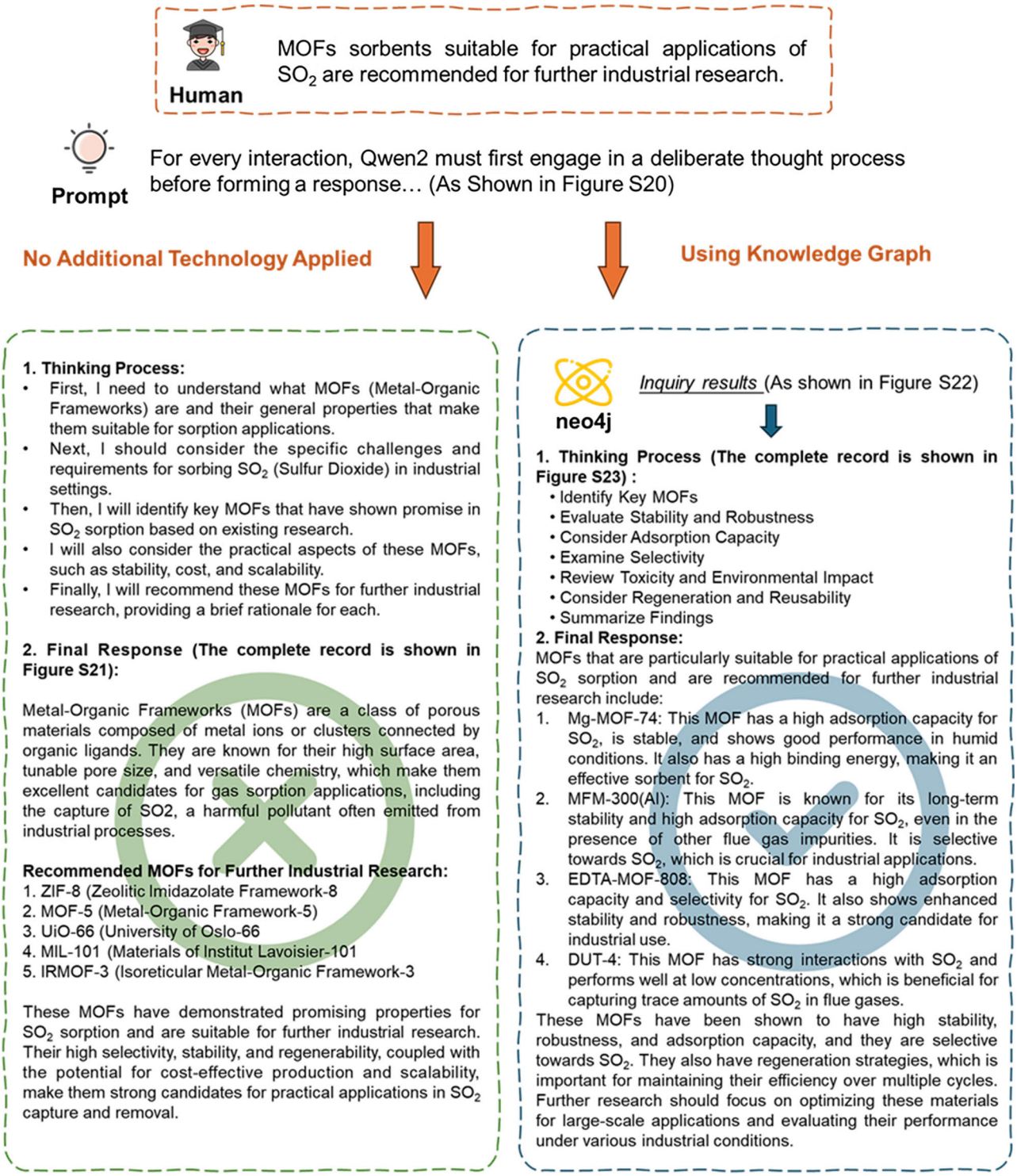

Fig. 6 | The CoT thinking process that enhances LLMs without and with knowledge graphs is used for industrial SO_2 adsorbent design.

model achieved an accuracy rate of 91.67 ± 0.94%, as assessed by the experts based on these criteria. This result further validates the significant advantage of our knowledge graph in enhancing the accuracy and effectiveness of answering questions within the domain of FMs. With the support of the knowledge graph, the Qwen2 model demonstrated a superior ability to comprehend and apply domain-specific knowledge, leading to the provision of more precise and insightful answers in materials science and related fields (Table S4).

CoT can effectively guide structured reasoning, helping models build answers step by step, which significantly advances the reasoning and innovation of LLMs in scientific problems. However, CoT techniques heavily rely on background knowledge. In the absence of sufficient background knowledge, even a well-structured reasoning process cannot yield good answers. Here, we attempt to introduce KG answers before CoT reasoning. As shown in the simple example in Fig. 6, we take the development of SO_2 adsorbents for industrial applications as an example and

prompt the LLMs to think in a CoT manner. Without using the KG, the LLMs, although clear in their reasoning, recommend materials with sub-optimal performance. In this case, LLMs (Prompt by Fig. S13) searched for relevant SO_2 information in the Knowledge Graph and passed it to the LLM (Prompt by Fig. S20). After analysis and reasoning (Figs. S21–S23), the LLM recommended a reasonable adsorbent material. In the future, integrating KG with CoT's thinking process may further enhance its ability for scientific discovery.

Discussion

In this study, we leveraged the exceptional natural language processing capabilities of LLMs to curate and examine over 100,000 articles. This extensive collection of textual data enabled us to construct a comprehensive knowledge graph, which encompasses an impressive 2.53 million nodes and 4.01 million relationships. Our research showcases the vast potential of this knowledge graph in enhancing data retrieval, facilitating data mining. and developing sophisticated question-answering systems in conjunction with LLMs. The enhanced LLMs, Qwen2-KG, empowered by a knowledge graph, can precisely respond to inquiries about FMs (accuracy rate 91.67%) and cite the sources of information. Incorporating KGs into the chain-of-thought reasoning of LLMs significantly enhances reasoning quality for a material screening task. The combined use of KGs and LLMs represents a pivotal step forward in AI for scientific research. It is expected that the comprehensive knowledge graphs of various research fields will be developed in the near future, to promote the automation and intelligence of scientific research through integration with AI tools.

Data availability

The data that supports the findings of this study are available at https://github.com/MontageBai/KGFM from the corresponding author upon reasonable request.

Code availability

The code supporting this study's findings is available at https://github.com/MontageBai/KGFM.

Received: 8 July 2024; Accepted: 22 January 2025;
Published online: 27 February 2025

References

1. Yaghi, O. M., Li, G. M. & Li, H. L. Selective binding and removal of guests in a microporous metal–organic framework. *Nature* **378**, 703–706 (1995).
2. Li, H., Eddaoudi, M., O'Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
3. Diercks, C. S. & Yaghi, O. M. The atom, the molecule, and the covalent organic framework. *Science* **355**, eaal1585 (2017).
4. Brunet, P., Simard, M. & Wuest, J. D. Molecular tectonics. porous hydrogen-bonded networks with unprecedented structural integrity. *J. Am. Chem. Soc.* **119**, 2737–2738 (1997).
5. Cerasale, D. J., Ward, D. C. & Easun, T. L. MOFs in the time domain. *Nat. Rev. Chem.* **6**, 9–30 (2021).
6. Choi, S., Kim, T., Ji, H., Lee, H. J. & Oh, M. Isotropic and anisotropic growth of metal–organic framework (MOF) on MOF: logical inference on MOF structure based on growth behavior and morphological feature. *J. Am. Chem. Soc.* **138**, 14434–14440 (2016).
7. Emmerling, S. T. et al. Interlayer interactions as design tool for large-pore COFs. *J. Am. Chem. Soc.* **143**, 15711–15722 (2021).
8. Gu, Y. et al. Controllable modular growth of hierarchical MOF-on-MOF architectures. *Angew. Chem. Int. Ed.* **56**, 15658 (2017).
9. Jain, C. et al. Tailoring COFs: transforming nonconducting 2D layered COF into a conducting quasi-3D architecture via interlayer knitting with polypyrrole. *J. Am. Chem. Soc.* **146**, 487–499 (2023).
10. Karmakar, A. et al. Hydrogen-bonded organic frameworks (HOFs): a new class of porous crystalline proton-conducting materials. *Angew. Chem. Int. Ed.* **55**, 10667 (2016).
11. Ming, X., Bin, L., Shilun, Q. & Banglin, C. Emerging functional chiral microporous materials: synthetic strategies and enantioselective separations. *Mater. Today* **19**, 503–515 (2016).
12. Peng, P. et al. Cost and potential of metal-organic frameworks for hydrogen back-up power supply. *Nat. Energy* **7**, 448–458 (2022).
13. Zick, M. E. et al. Fluoroarene separations in metal–organic frameworks with two proximal Mg2+ coordination sites. *J. Am. chem. Soc.* **143**, 1948–1958 (2021).
14. Wang, B., Lin, R.-B., Zhang, Z., Xiang, S. & Chen, B. Hydrogen-bonded organic frameworks as a tunable platform for functional materials. *J. Am. chem. Soc* **142**, 14399–14416 (2020).
15. Zhang, X. et al. Optimization of the pore structures of MOFs for record high hydrogen volumetric working capacity. *Adv. Mater.* **32**, 6 (2020).
16. Zhang, X., Li, Y. & Li, J.-R. Metal-organic frameworks for multicomponent gas separation. *Trends Chem.* **6**, 22–36 (2023).
17. Lv, X.-L. et al. A base-resistant metalloporphyrin metal–organic framework for C–H bond halogenation. *J. Am. Chem. Soc.* **139**, 211–217 (2017).
18. Yang, F. et al. A flexible metal–organic framework with a high density of sulfonic acid sites for proton conduction. *Nat. Energy* **2**, 877–883 (2017).
19. Lázaro, I. A., Wells, C. J. R. & Forgan, R. S. Multivariate modulation of the Zr MOF UiO-66 for defect-controlled combination anticancer drug delivery. *Angew. Chem. Int. Ed* **59**, 5211–5217 (2020).
20. Wang, P.-L. et al. Metal–organic frameworks for food safety. *Chem. Rev.* **119**, 10638–10690 (2019).
21. Xie, L.-H., Liu, X.-M., He, T. & Li, J.-R. Metal-organic frameworks for the capture of trace aromatic volatile organic compounds. *Chem* **4**, 1911–1927 (2018).
22. Zhang, H. et al. Charge and mass transport mechanisms in two-dimensional covalent organic frameworks (2D COFs) for electrochemical energy storage devices. *Energy Environ. Sci.* **16**, 889–951 (2023).
23. Fang, Y. et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nat. Mach. Intell.* **5**, 542–553 (2023).
24. Jiang, X., Zhu, R., Ji, P. & Li, S. Co-Embedding of Nodes and Edges With Graph Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 7075–7086 (2023).
25. Peng, Y., Au-Yong, C. P. & Myeda, N. E. Knowledge graph of building information modelling (BIM) for facilities management (FM). *Autom. Constr.* **165**, 105492 (2024).
26. Wang, C., Yang, Y., Song, J. & Nan, X. Research progresses and applications of knowledge graph embedding technique in chemistry. *J. Chem. Inf. Model.* **64**, 7189–7213 (2024).
27. Zheng, X., Wang, B., Zhao, Y., Mao, S. & Tang, Y. A knowledge graph method for hazardous chemical management: ontology design and entity identification. *Neural. Comput.* **430**, 104–111 (2021).
28. Zeng, X., Tu, X., Liu, Y., Fu, X. & Su, Y. Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* **72**, 114–126 (2022).
29. Rydholm, E. et al. Expanding the chemical space using a chemical reaction knowledge graph. *Digit. Discov.* **3**, 1378–1388 (2024).
30. Gao, Y., Wang, L., Chen, X., Du, Y. & Wang, B. Revisiting electrocatalyst design by a knowledge graph of Cu-based catalysts for CO_2 reduction. *ACS Catal* **13**, 8525–8534 (2023).
31. Yang, L., Chen, H., Li, Z., Ding, X. & Wu, X. Give us the facts: enhancing large language models with knowledge graphs for fact-aware language modeling. *IEEE Trans. Know.* **36**, 3091–3110 (2024).
32. Zhou, B. et al. CausalKGPT: industrial structure causal knowledge-enhanced large language model for cause analysis of quality problems in aerospace product manufacturing. *Adv. Eng. Inform.* **52**, 101611 (2024).

33. Yuan, A. et al. Knowledge graph question answering for materials science (KGQA4MAT). In *Research Conference on Metadata and Semantics Research*. 18–29 (2023).
34. Yuan, A. et al. Building open knowledge graph for metal-organic frameworks (MOF-KG): challenges and case studies. In *28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (2022).
35. Peng, C., Xia, F., Naseriparsa, M. & Osborne, F. Knowledge graphs: opportunities and challenges. *Artif. Intell. Rev.* **56**, 13071–13102 (2023).
36. Weis, J. W. & Jacobson, J. M. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nat. Biotechnol.* **39**, 1300–1307 (2021).
37. Zhong, L., Wu, J., Li, Q., Peng, H. & Wu, X. A comprehensive survey on automatic knowledge graph construction. *ACM Comput. Surveys* **56**, 1–62 (2023).
38. Venugopal, V. & Olivetti, E. MatKG: an autonomously generated knowledge graph in material science. *Sci. Data* **11**, 217 (2024).
39. Li, Y. & Starly, B. Building a knowledge graph to enrich ChatGPT responses in manufacturing service discovery. *J. Ind. Inf. Integr.* **40**, 100612 (2024).
40. Feng, F. et al. GenomicKB: a knowledge graph for the human genome. *Nucleic Acids Res.* **51**, D950–D956 (2022).
41. He, T. et al. Trace removal of benzene vapour using double-walled metal–dipyrazolate frameworks. *Nat. Mater.* **21**, 689–695 (2022).
42. Bai, X., Xie, Y., Zhang, X., Han, H. & Li, J.-R. Evaluation of open-source large language models for metal–organic frameworks research. *J. Chem. Inf. Model.* **64**, 4958–4965 (2024).

Acknowledgements

We acknowledge the financial support from the National Key Research and Development Program of China (2021YFB3501501), the National Natural Science Foundation of China (Nos. 22225803, 22038001, 22278011, and 22108007) and Beijing Natural Science Foundation (No. Z230023).

Author contributions

J.-R.L. and X.B. conceived and designed the study. X.B. wrote the initial manuscript. S.H., Y.L., and Y.X. participated in discussing and editing the manuscripts. J.-R.L., W.D., and X.Z. performed the supervision, review, and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-025-01540-6.

Correspondence and requests for materials should be addressed to Wenli Du or Jian-Rong Li.