RESEARCH ARTICLE

National Science Open
3: 20230040, 2024
https://doi.org/10.1360/nso/20230040

Chemistry

Special Topic: AI for Chemistry

Semantic knowledge graph as a companion for catalyst recommendation

Zhiying Zhang[1,#], Shengming Ma[1,#], Shisheng Zheng[1,#], Zhiwei Nie[1], Bingxu Wang[1], Kai Lei[2,*], Shunning Li[1,*] & Feng Pan[1,*]

[1]*School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, China;*
[2]*Shenzhen Key Lab for Information Centric Networking and Blockchain Technology (ICNLAB), Peking University Shenzhen Graduate School, Shenzhen 518055, China*

#Contributed equally to this work.
*Corresponding authors (emails: leik@pkusz.edu.cn (Kai Lei); lisn@pku.edu.cn (Shunning Li); panfeng@pkusz.edu.cn (Feng Pan))

Received 3 July 2023; Revised 17 November 2023; Accepted 17 January 2024; Published online 6 February 2024

Abstract: Our ability to perceive the correlation of different substances in the world is one of the key aspects of human intelligence. The passing of this faculty to artificial intelligence (AI) represents arguably one of the long-standing challenges in the application of AI to scientific problems. To meet this challenge in the burgeoning field of AI for chemistry, we may adopt the paradigm of knowledge graph. Herein, focusing on catalytic chemical reactions, we have developed a semantic knowledge graph framework based on both structured and unstructured data, the latter of which are extracted from the text of 220,000 articles on catalysts for organic molecules. The framework captures the latent knowledge of reactant-catalyst-product relationships and can therefore provide accurate recommendation on potential catalysts for targeted reaction, which especially facilitates the research involving large molecules. This study presents a viable pathway towards the implementation of literature-based data management in a catalyst recommendation platform.

Keywords: knowledge graph, text mining, catalysts, organic molecules, natural language processing

INTRODUCTION

Most of the state-of-the-art knowledge in scientific fields is scattered in the massive body of academic publications, generally without quantitative interpretation. This has prompted a need to extract and collate the valuable information from scientific literature, which could facilitate data-driven innovation and the discovery of latent knowledge. With the rapid development of artificial intelligence (AI) [1–3], atural language processing (NLP) [4–7] technology in company with domain-specific knowledge sources has sprung up and served as a companion to assist researchers in literature-based data mining. These AI steering tools can overcome the limitations of human processing speed, enabling the simultaneous analysis of billions of data points and offering substantial time saving. Among the typical NLP techniques, a knowledge graph [8–11] is a promising data management method that can provide a flexible and structured framework to organize information in diverse formats. In a knowledge graph, human knowledge is represented in the form of

entities, relationships, and semantic descriptions. Entities refer to objects and abstract concepts, while relationships depict connections between entities with types and attributes that have clear relevance to the corresponding field. Here in this work, we propose a procedure to construct the catalyst-specific knowledge graph in the field of chemical science, which can collaborate with scientists by recommending potential catalysts for the targeted reactants and/or products. This protocol is extendable to other relationship-prediction problems in multiple disciplines, which may benefit interdisciplinary research in chemistry, physics, biology, medicine, etc. [12–16].

Previous studies utilizing knowledge graph in science generally relied on a single data source [17–20]. For example, Mrdjenovich and co-workers [21] have established a knowledge graph of materials science using Materials Project data, with a focus on the estimation of physical properties through a graphical representation of the relationships among different properties. Recently, our group [22] has developed a semantic knowledge graph based on the abstracts of the articles for lithium-ion battery cathodes, which could tease out the text-mining processes and help identify potential cathode materials. However, in the field of catalysis [23–28], multi-source data fusion is presumably a prerequisite for the knowledge graph, because we need both a structured database of reliable catalytic relationships (reactant-catalyst-product) to serve as the backbone in the construction process and an unstructured database of academic texts so as to keep up to date with current findings in scientific literature.

In this regard, we propose a multi-source semantic knowledge graph framework that incorporates both structured and unstructured data, which is apt to handle relationship-prediction problems in the field of catalysis for organic molecules. The as-constructed catalytic knowledge graph (CatKG) functions as a convenient assistant for organizing and merging the domain knowledge extracted from the text of 220,000 research articles correlated with catalysts. We also demonstrate the capability of CatKG in learning and exploiting the underlying trends of the catalytic chemical reactions, which gives accurate recommendation of potential catalysts and points the way towards an AI-accelerated platform for chemistry research.

METHODS

Schema design of CatKG

We built the schema of CatKG based on the property graph as displayed in Figure S1. We started by establishing substance nodes for reactants, products, and catalysts, and then we extracted the element composition of each substance node and established a composition relation with the element node. The reactant-product relation edge (denoted as “Produce”) was established from the reactant to the product, along with the reactant-catalyst relation edge (denoted as “Catalyze”) from the reactant to the catalyst. Two types of nodes and three types of edges were thus abstracted, and the nodes of the same type were merged together to obtain the schema. Meanwhile, we added attributes to edges and nodes to enrich their semantics. For element nodes, we added attributes like group number and atomic number; for the substance node, we considered its different expressions, such as its name and SMILES string, and different physical properties, such as boiling point; for the reactant-product relation edge, we recorded the corresponding catalyst. In this schema, “catalyst” property was added to the reaction type, and multiple react edges were established for different catalyst properties between two nodes. For the theoretically predicted reaction, the reaction energy was also recorded.

Overall, the CatKG schema summarizes the key reactions and catalytic information that we care about in the field of catalysis in a concise form. The attributes of element nodes and substance nodes are displayed in Table S1.

Word2vec model

Before the training of word2vec model, we preprocessed the 220,000 articles collected with the keyword "catalyst". First, all contents after the "Conclusion" heading of each article were deleted, because they usually contain acknowledgements that are not relevant to the research. Then, the entire corpus was sequentially subjected to punctuation removal, tokenization, stop-word removal, and lemmatization, as shown in Figure S2. In this process, each sentence was divided into words by spaces, called tokens, and word2vec would analyze each input sentence in token units. Words like "this" and similar words that appear in abundance but have no real meaning were deleted as stop words. Lemmatization was performed to reduce the number of different forms of synonyms; for example, the third-person singular of the verb and the plural form of the noun were reduced to their prototypes. The above pre-processed corpus was used to train the word2vec model, which uses the interface of the gensim library, a python library for statistical NLP. The training parameters of the model are shown in Table S2. After training, the word2vec model was able to encode words into word vectors according to the semantics of the context. As shown in Tables S3 and S4, words with high or low similarity to a targeted word (such as "TiO_2") can be identified by the cosine similarity value, which quantifies the internal similarity between any two word-vectors. We identified words with high similarity to "TiO_2", such as "titania" (the English name of TiO_2), "anatase", "rutile", and "brookite" (the different phases of TiO_2). This result indicates the ability of word2vec model in representing the semantics of information in academic articles.

To extract catalysts from the word2vec model, we first used regular expressions to identify chemical formula entities from the literature, which can serve as a portfolio of candidate catalysts. Then, the cosine similarity of the chemical formula to the word vector of "catalyst" (or its similar words) was used to determine whether it is a catalyst or not:

$$score(X) = \max_{w \in V} \{cosSim(X, w)\} \tag{1}$$

$$\begin{aligned} V = \{&\text{“catalyst”}, \text{“catalytic”}, \text{“catalyze”}, \text{“catalysis”}, \text{“catalytically”}, \\ &\text{“catalyse”}, \text{“catalyzed”}, \text{“catalyzing”}, \text{“catalysts”}, \text{“catalytical”}, \\ &\text{“catalyzer”}, \text{“catalysed”}, \text{“catalyzation”}, \text{“catalysing”}\} \end{aligned} \tag{2}$$

where X represents the chemical formula obtained with a regular expression, V represents the catalyst synonym and cos*Sim* represents the word-vector similarity. After ranking the chemical formulas according to the above score from the highest to the lowest, the top 20 of them are displayed in Figure S3. We selected the chemical formulas with scores above 0.25 for the catalyst list, which corresponds to a total of 987 materials.

Advantages of the full-text corpus

The previous NLP models in scientific domains were generally based on the abstracts of articles. Here, we extracted the abstracts of the 220,000 articles and compared the model performance with that trained with the

full-text corpus. In Figure S4, the number of catalysts identified from different corpora as a function of the similarity score thresholds is shown. It turns out that the word2vec model trained on full-text corpus can uncover more catalysts than that on abstract corpus. For example, at a threshold value of 0.25, the model trained on full-text corpus output 987 catalysts, while that trained on abstract output 362 catalysts. We also found that the selection of thresholds significantly impacts the accuracy and robustness of the model. Opting for a lower threshold yielded a greater number of catalysts but led to a decrease in accuracy, while a higher threshold resulted in fewer extracted catalysts and potentially reduced interference from dirty data. In this work, we considered a threshold value (0.25) corresponding to the median number of catalysts that compromises with accuracy.

BERT model

The catalytic reactions in CatKG were used to train the BERT model. First, each catalytic reaction in CatKG was modeled as a "reactant-catalyst-product" sequence. To better characterize the reactants and products, they were represented by SMILES, while the catalyst was represented by a chemical formula. Reactants and products were tokenized at the character level. Afterwards, we used the span mask to randomly mask all tokens of the reactants, products, or catalysts. The corresponding token was replaced by a fixed <mask> tokens. The sequence was fed into the BERT model, which was asked to predict the original substance at the <mask> label. The training hyper-parameters of the BERT model are given in Table S5. In the encoding phase, the expression of the substance rather than the entire reaction sequence was fed into the model. The hidden layer states of the model were average-pooled to obtain a vector representation of the input. Each substance would be encoded as a 768-dimensional vector. To visualize these vectors in two dimensions, we use the t-SNE method for dimensionality reduction.

Catalyst prediction task

The predictive ability of CatKG can be manifested in two aspects: analogical inference using text information and reasoning using knowledge coding model. Analogical inference was carried out based on the existing catalytic relationships:

$$v(\text{catalyst}_1) - v(r_1) \approx v(\text{catalyst}_2) - v(r_2) \tag{3}$$

where $v(x)$ represents the word embedding of word x. This equation represents the translation invariance of word embeddings. The known catalytic relationship was filled into the left side of the formula, and the targeted reactant that we want to recommend catalyst for was set as r_2. Appropriate catalysts would be recommended from the database by searching for the highest cosine similarity between the left and right sides of the formula. The reasoning using knowledge coding model corresponds to the catalyst prediction task from BERT model. The model assigned a score to all tokens, and the three materials with the highest score were considered the predicted catalysts for the targeted reaction. That is to say, the model will recommend three catalysts in the prediction, and in this work, if any of the recommended catalysts appeared in the test set, the recommendation would be considered correct.

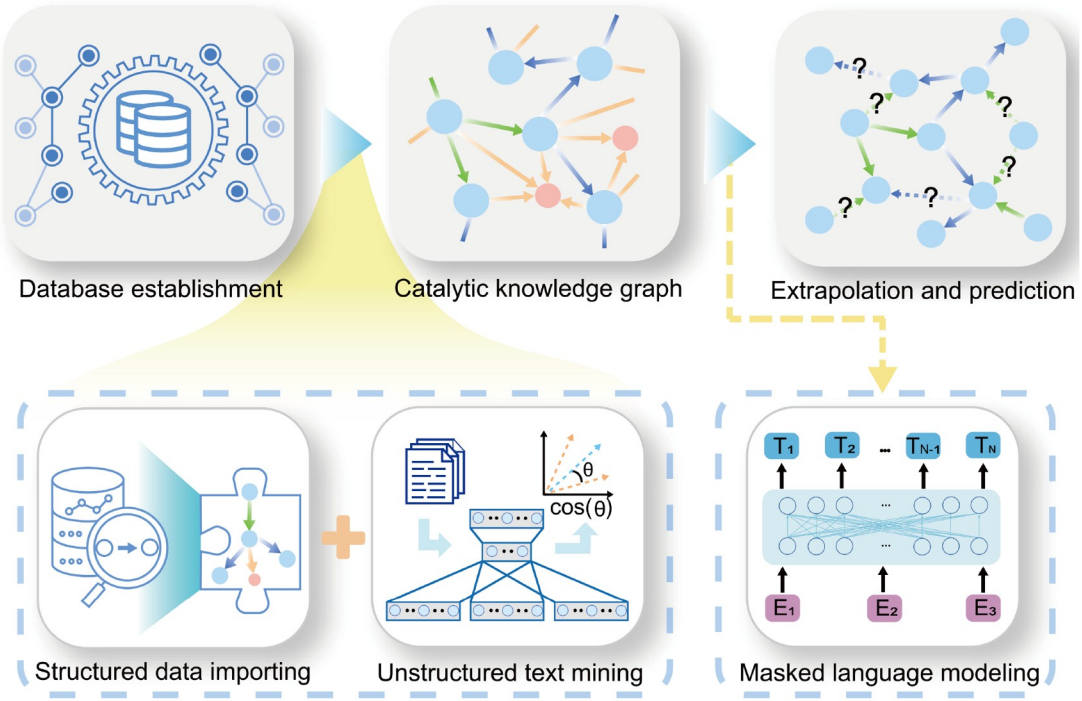

Figure 1 Conceptual workflow of text mining using a semantic knowledge graph. This protocol can produce natural language documentation for both structured and unstructured data in the domain of catalytic chemical reactions, and is predictive of catalytic relationships.

RESULTS AND DISCUSSION

A schematic workflow of text mining using a multi-source semantic knowledge graph is illustrated in Figure 1. To begin with, both structured and unstructured data were collected, from which we can design the schema format that defines the types of nodes and edges in the knowledge graph (see METHODS for details). The nodes consist of two types, including substance and element; while the edges consist of three types, including composition relation, reactant-product relation, and reactant-catalyst relation. A property graph model was employed to represent data, allowing nodes and edges to have their own attribute lists. Following this procedure, we developed a word embedding module by utilizing the unstructured text. The module was then integrated with the structured data to form the knowledge graph according to the schema. In addition, a knowledge encoding module using the masked language modeling task was introduced, which is central to the intelligent capability of predicting new catalytic relationships. By navigating these relationships, we may identify the potential catalyst that, although it exists in the literature, is still unexplored for a targeted reaction we are interested in. This pipeline works as a catalyst recommendation system without any expert scrutiny.

The detailed architecture of CatKG is shown in Figure 2. The corpus of structured data was retrieved from the existing databases of organic molecules and their corresponding catalysts in a well-formatted manner [28,29]. An ontology extractor was implemented to extract the critical information from this corpus, which generates the initial nodes and edges of CatKG. For the corpus of unstructured text data, both article abstract and full text were employed to train the word2vec model, which encodes words into vectors according to the contextual semantic relations. A total of 220,000 research articles containing the keyword “catalyst” were

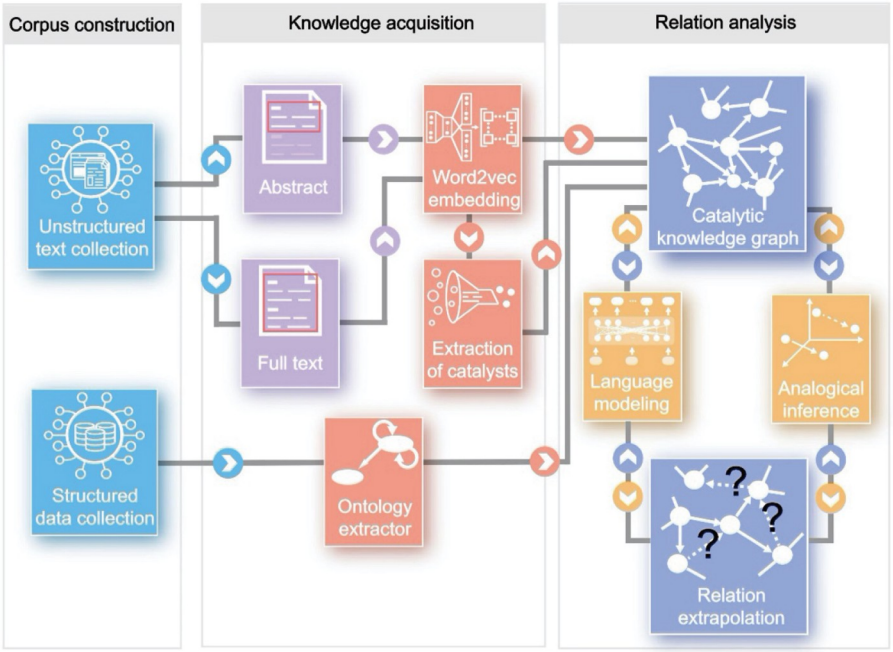

Figure 2 Architecture of CatKG. Using both structured databases and unstructured text sources as the corpora. CatKG is constructed on an ensemble of reactant-catalyst-product relationships, which can enable automatic catalyst recommendation via extrapolation of word embeddings.

adopted for this embedding model, and entities with a high word-vector similarity to "catalyst" (or related words) would be added to the nodes of substance after a simple rule-based classification (Figure S3). Through this step, we could feed the knowledge graph with text data from contemporary research findings. Furthermore, by comparing the outputs with and without utilizing full text for training, we illuminate that the number of extracted catalysts was considerably higher when a full-text corpus was incorporated (Figure S4). This underscores the significance of full text in augmenting the search space for potential catalysts based on CatKG. As a final step, the reasoning capability of CatKG is accommodated via a combination of the popular language model BERT [30] and the analogical inference of word embeddings of existing catalysts and reactants (see METHODS for details). New catalytic relationships will be inferred from these relation extrapolation schemes, thus allowing for the exploration of catalysts for various chemical reactions.

An exemplified set of nodes and edges in CatKG are visualized in Figure 3A, justifying the successful establishment of reactant-catalyst-product relationships. These relationships, along with more specific reaction conditions, can be stored as attributes for the reactant-product relation edges, thereby achieving high flexibility and scalability. We would like to emphasize that node- and edge-centric queries could be performed in the network to locate the reaction path as needed. The real chemical space is much larger than displayed here, since a total of 11,649 catalytic reactions have been employed as the dataset for the BERT model. To visualize this, we make a projection of the high-dimensional descriptors for reactants and products onto two dimensions, as shown in Figure 3B. This mapping can enable an evaluation of the model efficiency in capturing the underlying features of the catalytic reactions. We take the reactants/products related to the Cu catalyst as an example, indicated by the red/blue dots in the left panel of Figure 3B. The distribution of most of these data points exhibits a noticeable tendency to be organized into multiple groups, as highlighted in the

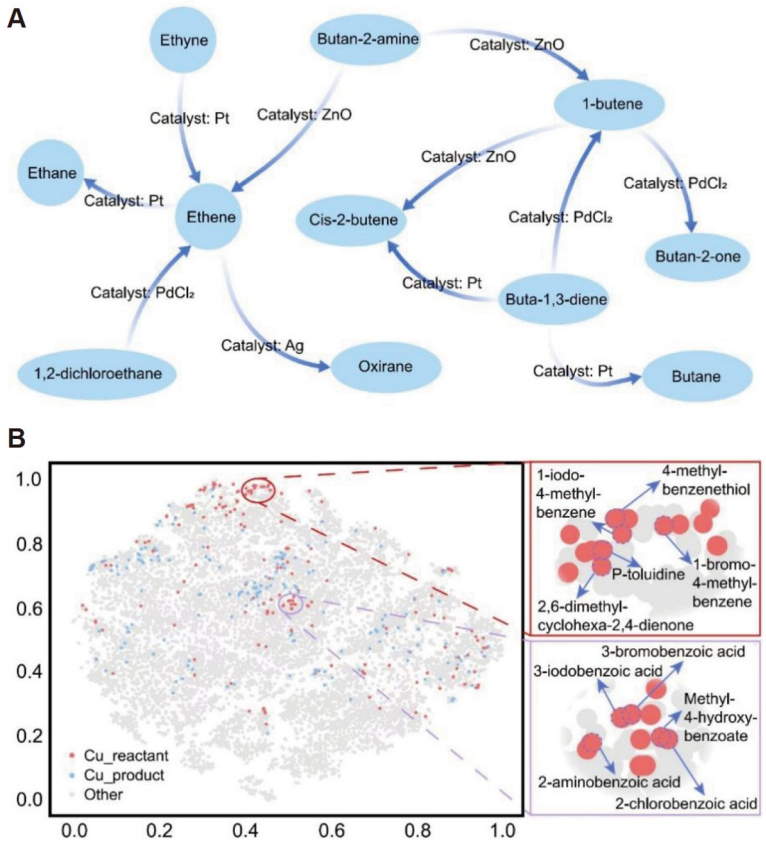

Figure 3 (A) A subset of nodes and edges in CatKG. (B) Projection of the BERT-encoded representations to a 2D map using t-distributed stochastic neighbor embedding (t-SNE) [31] method.

right panels. Data points in close vicinity generally imply that they have similar catalytic relationships, in line with their common correlation with the Cu catalyst. Yet, the fact that they do not group into a single cluster indicates a sufficient diversity of the reactants/products in the database.

The capability for predicting potential catalytic relationships by CatKG can be demonstrated by the example of Pt and ethylene (Figure 4A), the former being a catalyst for the reaction (either redox or isomerization) of the latter. By relation extrapolation in CatKG, it is predicted that Pt can also catalyze the reactions for propylene and butene. Further extrapolation leads to the identification of TiO_2 as a catalyst for these two molecules, and the identification of RuO_2 for catalyzing the reaction of butene. Some of the above results have been justified by previous studies (e.g., the pair of Pt and propylene), while others warrant further experimental examination.

Moreover, as an intriguing research topic, the prediction of catalysts for large molecules from the catalytic relationships of small molecules could be a fundamental contribution to chemical science. Here, we test whether CatKG can effectively accomplish this job. Specifically, we selected five different cutoff values ranging between 100 and 500 for the relative molecular masses of the molecules as either reactants or products (we note that the relative molecular masses are below 2000 for all molecules in this work). In the training set, we included molecules with relative molecular masses below the cutoff value, while in the test

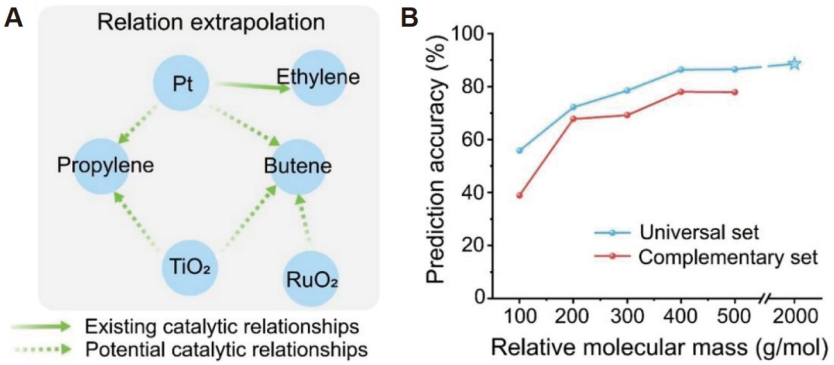

Figure 4 (A) The predictive ability of CatKG, using the Pt-ethylene pair as a basis for deduction. (B) Results of predicting the catalytic relationships of large molecules using small molecules as the training set. The horizontal coordinate corresponds to the threshold of relative molecular mass, and only below this threshold can the molecule be included in the training set as a reactant/product. The red line denotes the results of a test set, in which only above the threshold as defined by the horizontal coordinate can the molecule be included. The blue line denotes the results without such threshold.

set, we included those above this value. Separate BERT models were trained for each of these thresholds, and the results of the test sets are depicted in Figure 4B as a red line. We note that the accuracy can reach about 80% if the cutoff value is set at 400 or above, and this is practical for use as a catalyst recommendation system. We also examined the case where the cutoff value was not applied to the test set; that is, molecules are included for testing regardless of their relative molecular masses (the blue line in Figure 4B). Notably, with higher cutoff values, the accuracy quickly approaches its maximum, as indicated by an asterisk, suggesting that the underlying trends of the catalytic chemical reactions can be directly learned from the relatively small molecules. This capability, when integrated with the prior knowledge of a researcher, can greatly accelerate our search for promising catalysts regarding the complicated chemical reactions of large organic molecules. Altogether, the CatKG model proposed in this work has proven the efficacy of a semantic knowledge graph for the automatic generation of potential catalytic relationships between entities in chemistry. We believe that future collaboration between knowledge graphs and large language models will afford the opportunity for a more convenient and accurate AI system across all domains in science.

CONCLUSION

In summary, we have described a multi-source semantic knowledge graph framework for text mining the information of catalytic relationships in chemistry. We demonstrated the success of CatKG in data management for the reactant-catalyst-product relationships extracted from scientific literature, which can capture the hidden trends of the chemical reactions among organic molecules. Such an ability could allow for catalyst recommendation with high accuracy, which will speed up scientific research related to complicated large molecules. Our protocol is readily applicable to the relationship-prediction tasks in other scientific domains, especially when both structured database and unstructured text information are indispensable for the AI-assisted decision making.

Funding

This work was supported by Guangdong Basic and Applied Basic Research Foundation (2023A1515011391 and 2020A1515110843), the Soft Science Research Project of Guangdong Province (2017B030301013), the National Key Research and Development Program of China (2022YFB2702301), the Key-Area Research and Development Program of Guangdong Province (2020B0101090003), and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen.

Author contributions

S.L. and F.P. proposed and supervised the project. Z.Z., S.M., and S.Z. conceived the model. Z.Z. contributed to algorithm design and analyzed the predicted results. S.M. conceptualized the schema and fine-tuned the algorithms. S.Z. contributed to the evaluation of the knowledge graph. Z.N. and B.W. collected the datasets. K.L. advised on the development of the model. S.L., Z.Z., S.M. and S.Z. wrote the manuscript. All authors participated in the discussion and agreed with the conclusions of the study.

Conflict of interest

The authors declare no conflict of interest.

Supplementary information

The supporting information is available online at https://doi.org/10.1360/nso/20230040. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Butler KT, Davies DW, Cartwright H, *et al.* Machine learning for molecular and materials science. *Nature* 2018; **559**: 547–555.
2 de Almeida AF, Moreira R, Rodrigues T. Synthetic organic chemistry driven by artificial intelligence. *Nat Rev Chem* 2019; **3**: 589–604.
3 Gomes CP, Selman B, Gregoire JM. Artificial intelligence for materials discovery. *MRS Bull* 2019; **44**: 538–544.
4 Pei Z, Yin J, Liaw PK, *et al.* Toward the design of ultrahigh-entropy alloys via mining six million texts. *Nat Commun* 2023; **14**: 54.
5 Kononova O, Huo H, He T, *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci Data* 2019; **6**: 203.
6 He T, Sun W, Huo H, *et al.* Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem Mater* 2020; **32**: 7861–7873.
7 Kumar A, Ganesh S, Gupta D, *et al.* A text mining framework for screening catalysts and critical process parameters from scientific literature—A study on hydrogen production from alcohol. *Chem Eng Res Des* 2022; **184**: 90–102.
8 Lin Y, Liu Z, Sun M, *et al.* Learning entity and relation embeddings for knowledge graph completion. In: Twenty-ninth AAAI Conference on Artificial Intelligence. Austin, 2015.
9 Pujara J, Miao H, Getoor L, *et al.* Knowledge graph identification. In: International Semantic Web Conference. Athens, 2013, 542–557.
10 Wang Q, Mao Z, Wang B, *et al.* Knowledge graph embedding: A survey of approaches and applications. *IEEE Trans Knowl Data Eng* 2017; **29**: 2724–2743.
11 Nie Z, Liu Y, Yang L, *et al.* Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of $LiFePO_4$. *Adv Energy Mater* 2021; **11**: 2003580.
12 Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language

processing: Interpreting hypernymic propositions in biomedical text. *J BioMed Inf* 2003; **36**: 462–477.

13 Rindflesch TC, Kilicoglu H, Fiszman M, *et al.* Semantic MEDLINE: An advanced information management application for biomedicine. *Inform Serv Use* 2011; **31**: 15–21.

14 Gu Y, Tinn R, Cheng H, *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc* 2022; **3**: 1–23.

15 Hong L, Lin J, Li S, *et al.* A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nat Mach Intell* 2020; **2**: 347–355.

16 Manica M, Mathis R, Cadow J, *et al.* Context-specific interaction networks from vector representation of words. *Nat Mach Intell* 2019; **1**: 181–190.

17 Harnoune A, Rhanoui M, Mikram M, *et al.* BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Comput Methods Programs Biomed Update* 2021; **1**: 100042.

18 Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J* 2020; **18**: 1414–1428.

19 Santos A, Colaço AR, Nielsen AB, *et al.* A knowledge graph to interpret clinical proteomics data. *Nat Biotechnol* 2022; **40**: 692–702.

20 Wang X, Meng L, Wang X, *et al.* The construction of environmental-policy-enterprise knowledge graph based on PTA model and PSA model. *Resour Conserv Recycl Adv* 2021; **12**: 200057.

21 Mrdjenovich D, Horton MK, Montoya JH, *et al.* Propnet: A knowledge graph for materials science. *Matter* 2020; **2**: 464–480.

22 Nie Z, Zheng S, Liu Y, *et al.* Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes. *Adv Funct Mater* 2022; **32**: 2201437.

23 Aramouni NAK, Touma JG, Tarboush BA, *et al.* Catalyst design for dry reforming of methane: Analysis review. *Renew Sustain Energy Rev* 2018; **82**: 2570–2585.

24 Guo W, Zhang K, Liang Z, *et al.* Electrochemical nitrogen fixation and utilization: Theories, advanced catalyst materials and system design. *Chem Soc Rev* 2019; **48**: 5658–5716.

25 Abdulrasheed A, Jalil AA, Gambo Y, *et al.* A review on catalyst development for dry reforming of methane to syngas: Recent advances. *Renew Sustain Energy Rev* 2019; **108**: 175–193.

26 Garg S, Li M, Weber AZ, *et al.* Advances and challenges in electrochemical CO_2 reduction processes: an engineering and design perspective looking beyond new catalyst materials. *J Mater Chem A* 2020; **8**: 1511–1544.

27 Feng X, Liu H, He C, *et al.* Synergistic effects and mechanism of a non-thermal plasma catalysis system in volatile organic compound removal: A review. *Catal Sci Technol* 2018; **8**: 936–954.

28 Winther KT, Hoffmann MJ, Boes JR, *et al.* Catalysis-Hub.org, an open electronic structure database for surface reactions. *Sci Data* 2019; **6**: 75.

29 Shanghai Institute of Organic Chemistry of CAS. Chemistry Database [1978–2023]. https://organchem.csdb.cn.

30 Devlin J, Chang MW, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. arxiv: 1810.04805, 2018.

31 Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**: 2579–2605.