Exploring Pre-Trained Language Models to Build Knowledge Graph for Metal-Organic Frameworks (MOFs)

Yuan An, Jane Greenberg, Xiaohua Hu
Alex Kalinowski, Xiao Fang
Xintong Zhao, Scott McCLellan
Information Science
Drexel University
Philadelphia, PA, USA
{ya45,jg3243,xh29}@drexel.edu

Fernando J. Uribe-Romo
Kyle Langlois
Jacob Furst
Department of Chemistry
University of Central Florida
Orlando, FL, USA
fernando@ucf.edu

Diego A. Gómez-Gualdrón
Fernando Fajardo-Rojas
Katherine Ardila
Chemical and Biological Engineering
Colorado School of Mines
Golden, CO, USA
dgomezgualdron@mines.edu

Semion K. Saikin
Kebotix, Inc.
Cambridge, MA, USA
semion@kebotix.com

Corey A. Harper
Ron Daniel Jr.
Elsevier Labs
New York, NY, USA
{c.harper,r.daniel}@elsevier.com

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020568

***Abstract*—Building a knowledge graph is a time-consuming and costly process which often applies complex natural language processing (NLP) methods for extracting knowledge graph triples from text corpora. Pre-trained large Language Models (PLM) have emerged as a crucial type of approach that provides readily available knowledge for a range of AI applications. However, it is unclear whether it is feasible to construct domain-specific knowledge graphs from PLMs. Motivated by the capacity of knowledge graphs to accelerate data-driven materials discovery, we explored a set of state-of-the-art pre-trained general-purpose and domain-specific l anguage m odels t o e xtract k nowledge t riples f or metal-organic frameworks (MOFs). We created a knowledge graph benchmark with 7 relations for 1248 published MOF synonyms. Our experimental results showed that domain-specific PLMs consistently outperformed the general-purpose PLMs for predicting MOF related triples. The overall benchmarking results, however, show that using the present PLMs to create domain-specific knowledge graphs is still far from being practical, motivating the need to develop more capable and knowledgeable pre-trained language models for particular applications in materials science.**

***Index Terms*—Knowledge Graph, Pre-trained Language Model, Prompt Probing, Materials Science, Metal-Organic Frameworks**

I. Introduction

Knowledge graphs (KG) are a hallmark for representing domain knowledge in a graph structure with edges being a set of triples in the format of $\langle \mathsf{head}, \mathsf{predicate}, \mathsf{tail} \rangle$. Each triple captures a relationship (the predicate) between a subject entity (the head) and an object entity (the tail). A domain knowledge graph provides an easy way to query and reason about domain knowledge. Despite this ease, building a knowledge graph is a time-consuming and costly process, given the aim to extract and organize information from a wide variety of sources including vast amount of unstructured texts. Extracting triples from textural sources requires complex, costly, and specific natural language processing methods [5], [35]. For example, in data-driven materials science, researchers constantly trudge through various journal articles, patents, or company reports to glean useful research and experimental results. The growing number of scientific publications and the wide variety of ways that scientists publish their findings have posed a significant challenge to building knowledge graphs [3], [54].

Pre-trained large **L**anguage **M**odels (PLM) such as BERT [16], RoBERTa [31], GPT-3 [10], and T5 [40] have emerged as a crucial type of approach that provides readily available knowledge to a range of AI applications [48]. PLMs have attracted a significant attention in AI and NLP communities. A recent emerging paradigm leveraging the PLMs is to use textual prompts to solve problems. For example, for Knowledge Graph Construction (KGC), given a piece of text "`HKUST-1 is a metal organic framework.`", we can use a textual prompt "`HKUST-1 is a metal organic framework. HKUST-1 contains __ which is a metal.`" to ask a PLM to fill up the blank with a chemical element such as `Cu` or `Copper`. The downstream applications using this paradigm are reformulated to predict a missing or next word using a PLM. We designate this paradigm as *prompting on pre-trained large language models* or ***prompt@PLM*** for short.

In a recent challenge for Knowledge Base Construction from Pre-trained Language Models (KBC-LM) [1] which is collocated

[1] https://lm-kbc.github.io/

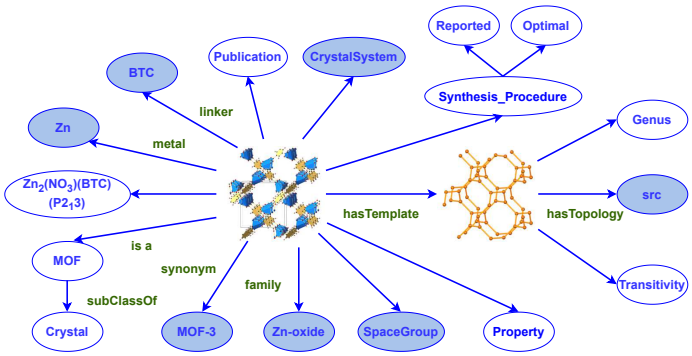

Fig. 1. The knowledge graph of metal-organic frameworks (MOF-KG) describes the structural, chemical, electric, and physical properties of MOFs that were gleaned from multiple disparate sources. In this study, we investigated the performance of a set of pre-trained language models for extracting the information of a MOF depicted in the shaded ovals of the figure.

with the 21st International Semantic Web Conference (ISWC-2022), participants constructed knowledge bases in general domains from a set of pre-trained language models including BERT-related LMs [16], RoBERTa, Transformer-XL, GPT-2, BART, etc. Various systems including our own [2], [19] have achieved the overall macro average recall ranging from 53% to 69%, overall macro average precision from 73% to 80%, and overall macro average F1-scores ranging from 49% to 67%, a significant improvement to the existing baseline LAMA (LAnguage Model Analysis) system [38] (ref. Table IV).

Inspired by this effort and the promising results of KBC-LM@ISWC-2022, we turn our attention to an under-explored area which aims to directly extract structured knowledge from PLMs for scientific domains, for example, materials science. A feasible approach of doing *prompt@PLM* for scientific domains would greatly reduce the cost and expedite the process of knowledge graph construction. Motivated by data-driven materials discovery, we created a knowledge graph benchmark for a particular type of material, metal-organic frameworks (MOF), and investigated a set of state-of-the-art general-purpose and domain-specific PLMs[2]. Figure 1 shows the ontological definition for the knowledge graph of metal-organic frameworks (MOF-KG). Our benchmark framework consists of triples described by the shaded entities and their associated relationships.

This rest of the paper reports the study and our findings. In particular, Section II discusses related work. Section III describes MOFs and the information captured by the MOF-KG. Section IV details the process of creating the MOF-KG benchmark. Section V reviews the PLMs we chose to explore. Section VI describes the relation-specific prompts for LM probing. Section VII outlines the experiments. Section VIII presents the results and findings. Finally, Section IX points to future work and concludes the paper.

[2]https://github.com/anyuanay/MOF_KG_LAMA

II. Related Work

An increasing number of studies have been reported using *prompt@PLM* to solve text classification [10], [20], [41], [42], named-entity recognition [14], natural language inference [20], [41], [42], sentiment analysis [28], relation extraction [12], [22], text summarization [1], and parsing [13]. For Knowledge Base Construction from LM Probing, the seminal work is LAMA (LAnguage Model Analysis) [38] which manually created cloze templates to probe knowledge in PLMs. Few-shot learning on the original LAMA datasets has also been evaluated [24]. More studies have been reported on probing PLMs for complicated knowledge [39], temporal knowledge [17], and domain specific knowledge such as BioLAMA [45] and MedLAMA [32]. However, applications of PLMs in materials science are just scarcely explored.

III. Metal-Organic Frameworks (MOF) Knowledge Graphs (KG)

In this section, we introduce the type of material, metal-organic frameworks. The underlined phrases in the following description highlight the concepts and associations we seek to capture in the knowledge graph. Metal-Organic Frameworks (MOFs) are a class of modular, porous crystalline materials that have great potential to revolutionize applications such as gas storage, molecular separations, chemical sensing, catalysis, and drug delivery. The crystal structures of MOFs can be (conceptually) modified by "swapping" constituent building blocks corresponding to metal-based clusters and organic linkers. These building blocks are interconnected in a pattern described by an underlying net. Figure 2 shows the framework structure of the MOF with the synonym 'MOF-3' and its underlying net which is coded as 'srs' using the RCSR database [37].

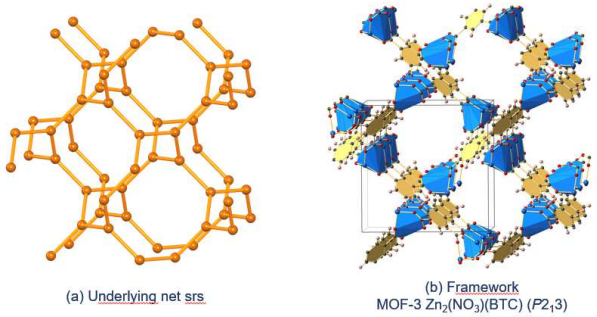

Fig. 2. The Underlying Net and Framework of MOF-3

The scientific potential of MOFs is primarily due to their usually high surface area and exceptionally tunable properties [50]. But the combinatorics of such building blocks means that chemists have access to a (not fully explored) "material design space" of trillions of structures. The sheer number has made the identification of optimal MOFs (and subsequent) synthesis for a given application a significantly challenging task. As a result, considerable effort have been put into developing

#	Relation	Description	Example
1	hasType	The published name (s) is a type of MOF material (o)	(MROF-1, hasType, [MOF, Metal-Organic Framework])
2	hasMetals	The named MOF (s) has metal clusters (o)	(HKUST-1, hasMetals, [Cu, Copper])
3	hasOrganicLinker	The named MOF (s) has organic linker (o)	(MOZIF-1, hasOrganicLinker, [O=N(=O)C1=NC=C[N]1])
4	hasMOFFamily	The named MOF (s) is in the MOF family (o)	(TMU-60, hasMOFFamily, [Zn-oxide, zinc oxide])
5	hasCrystalSystem	The named MOF (s) has a crystal system (o)	(URMOF-4, hasCrystalSystem, [trigonal])
6	hasTopology	The named MOF (s) has a topology code (o)	(IRMOF-61, hasTopology, [pcu])
7	hasSpaceGroup	The named MOF (s) has a space group (o)	(NTU-5, hasSpaceGroup, [C2/c])

TABLE I
MOF-KG BENCHMARK RELATION NAMES, DESCRIPTIONS, AND EXAMPLE TRIPLES. IN THE DESCRIPTION, (S) INDICATES SUBJECTENTITY AND (O) FOR OBJECTENTITY.

effective computational techniques to screen and isolate candidate MOF structures for the application of choice. Previous efforts include the creation of large MOF databases which contain both synthesized and "hypothesized" MOF structures [4]. However, a large amount of critical information about MOF properties and synthesis procedures remains scattered in scientific literature. Here, it is simply impossible for a human to scan and glean relevant information [49]. A MOF knowledge graph that extracts and integrates data from both MOF databases and scholarly articles presents a novel approach to identifying MOF prediction, discovery, and synthesis.

A MOF-KG requires an ontology that defines member terms. Due to the rapid development in the related field, there is not a general agreed system of nomenclature for describing MOFs and associated activities. Prior initiatives have attempted to standardize terminologies [7], [37], however, the diversity in the focus and the scientific inquiry can lead to a variety of terminological usages for this class of compounds. For the purpose of building the MOF-KG, we analyzed the structural and chemical information of many MOFs in the Cambridge Structural Database (CSD) [33] and propose an MOF ontology as illustrated in Figure 1. In the next section, we describe the process of generating triples for the knowledge graph benchmark.

IV. BUILDING THE MOF-KG BENCHMARK

We selected a set of 7 relations, each covering an aspect of MOF. For each relation, we generated a set of $\langle \mathsf{SubjectEntity}, \mathsf{relation}, \mathsf{ObjectEntity} \rangle$ triples as ground truth. Table I lists the relations along with their descriptions and ground truth examples.

We generated the benchmark data based on the MOF collection [33] in the Cambridge Structural Database (CSD) curated by the Cambridge Crystallographic Data Centre (CCDC), a world-leading organization that compiles and maintains small-molecule organic and metal-organic crystal structures. The CSD MOF collection contains approximately 16,300 successful MOFs structures (crystals) that have been realized experimentally, with crystal structure measured and solved for using diffraction techniques (X-rays, neutrons, electrons). We first queried the CSD MOF collection to extract 1,248 MOFs with published synonyms. We use these synonyms as MOF names for further data collection as illustrated below. The underlined phrases correspond to the types of triples captured in the benchmark.

- The 1,248 names are considered as a type of MOF (i.e, Metal-Organic Framework) material.
- We queried the CSD crystal database to extract the information about crystal system and space group.
- We used the CSD ConQuest tool to identify a MOF's family by applying the search criteria developed by Moghadam et al. in [34]. There are six prototypical MOF families identified: Zr-oxide nodes (e.g. UiO-66), Cu–Cu paddlewheels (e.g. HKUST-1), ZIF-like, Zn-oxide nodes, IRMOF-like, and MOF-74/CPO-27-like materials.
- Finally, we leveraged the MOFid system[3] developed by Bucior et al. in [11] to identify a MOF's metals, organic linker, and topology.

After building the MOF-KG benchmark, we probed the pre-trained language models to measure their capabilities for constructing the knowledge graph.

V. PRE-TRAINED LANGUAGE MODELS

General Architecture. Standard language models are trained to predict text in an autoregressive fashion, that is, predicting

[3]https://github.com/snurr-group/mofid

the tokens in the sequence one at a time. The sequence generally progresses from left to right, but alternative sequences can also be pursued. Representative examples of modern pre-trained left-to-right autoregressive LMs include GPT-3 [10]. A disadvantage of the autoregressive language models lies in the directionality of processing text. To predict text based on surrounding text, masked language models (MLM) have been developed that use bidirectional objective function. Representative pre-trained models using MLMs include BERT [16], ERNIE [53] and many variants. An alternative class is the prefix LM, a left-to-right LM that decodes a target text y conditioned on a prefixed sequence x as for translation. Example prefix LMs include UniLM 1-2 [6], [18] and ERNIE-M [36]. Another approach is the encoder-decoder model which uses a left-to-right LM to decode a target text y conditioned on a separate encoder for text x with a fully-connected mask. Example encoder-decoder pre-trained models include T5 [40], BART [27], MASS [44] and their variants.

Chosen PLMs for this Study. For triple extraction, we focus on Masked Language Models (MLM) given that they are trained to predict text by surrounding context. We aim to probe both general-purpose and domain-specific PLMs for constructing the MOF-KG. We selected two domain-specific BERT-like models trained on materials science corpora, MatBERT [46] and MatSciBERT [21]. We also selected the SciBERT PLM [8] trained on general science related text. For general-purpose PLMs, we chose the BERT large-cased PLM and an optimized variation, RoBERTa [31]. Table II lists the characteristics of the chosen PLMs for this study.

PLM	Size	Training Corpora
BERT-large	340M parameters	BooksCorpus (800M tokens) and English Wikipedia (2.5B tokens)
RoBERTa	355M parameters	BooksCorpus (800M tokens) and English Wikipedia (2.5B tokens) CC-NEWS (63M news articles) OPENWEBTEXT (36GB web content) STORIES (31GB story text)
SciBERT	similar to BERT	1.14M scientific papers from Semantic Scholar (3.1 billion tokens)
MatBERT	similar to BERT	two million peer-reviewed materials science journal articles (61 million paragraphs, 8.8B tokens)
MatSciBERT	similar to SciBERT	continue pre-training SciBERT with 150K papers from Elsevier Science Direct Database spanning four materials science families: inorganic glasses, metallic glasses, alloys, and cement and concrete

TABLE II
CHARACTERISTICS OF THE CHOSEN PLMS

VI. PROMPT DESIGN FOR PROBING THE PLMS

In general, there are two types of prompts. Cloze prompts are those which fill in blanks in a textual string. Prefix prompts differ in that they continue filling a prefixed string, rather than just a blank token. Prompts can be designed manually based on human intuition [10], [38], [41] or automatically through mining [26], paraphrasing [23], [52], gradient-based search [47], generation [9], and scoring [15]. In addition to discrete hard prompts, researchers have also developed continuous soft prompts that interact directly with LMs in the embedding space. Soft prompts have their own parameters that can be tuned through different strategies including prefix tuning [29], hard-prompt initialized tuning [55], and hybrid tuning [30].

1.	`[SUB]` contains `[MASK]` which is a metal.
2.	`[SUB]` contains a metal which is `[MASK]`.
3.	`[SUB]` is a metal organic framework. `[SUB]` contains `[MASK]` which is a metal.
4.	As a metal organic framework, `[SUB]` contains a metal which is `[MASK]`.
5.	`[SUB]` is a MOF. `[SUB]` has SBU metal `[MASK]`.
6.	`[SUB]` is MOF. `[SUB]` contains `[MASK]`.
7.	`[SUB]` contains `[MASK]`.
8.	`[SUB]` is a metal organic framework. `[SUB]` contains `[MASK]`.
9.	The SBU of `[SUB]` is `[MASK]`.
10.	`[SUB]` is a metal organic framework structure composed of metal cluster `[MASK]`.
11.	`[SUB]` is a metal organic framework. `[SUB]` has SBU metal `[MASK]`.
12.	`[MASK]` was used as metal center in the synthesized `[SUB]` material.

TABLE III
THE TEMPLATES FOR GENERATING PROMPTS FOR THE RELATION "HASMETALS"

In this study, we focus on manually developing templates for generating prompts. A template modifies the original text by adding extra tokens. For example, the template "`[SUB] is a metal organic framework. [SUB] contains [MASK] which is a metal.`" generates the prompt we used for Knowledge Graph Construction (KGC), where "`[SUB]`" corresponds to the SubjectEntity in the original text, and the token "`[MASK]`" stands for a blanked-out ObjectEntity to be filled up. In our case, the original text is a triple such as "⟨`HKUST-1, hasMetals, Copper`⟩". For each relation, we crafted 12 different templates for generating prompts. Each template is designed carefully based on analyzing the relevant publications that report some MOFs' structural and property information. Table III displays the 12 templates for the relation hasMetals. The templates designed for other relations are available in the Jupyter notebooks containing experiments in the github repository here (link provided in Footnote 2).

Given a triple ⟨SubjectEntity, predicate, ObjectEntity⟩ in the benchmark dataset, a template will generate a prompt by replacing `[SUB]` with the SubjectEntity. The generated prompt is sent to a PLM to predict the ObjectEntity by filling up the `[MASK]` in the prompt. In next section, we describe the experimental process and outline how to evaluate the answers returned by a PLM.

VII. EXPERIMENTS AND EVALUATION

We probed the 5 selected PLMs (Table II) using 7 relations (Table I) from the MOF-KG. Each relation is instantiated as a number of triples. Due to the limits of the data collection process (Section IV), not all the 1,248 named MOFs have complete number of triples for all relations. The minimum

number of triples for a relation (hasFamily) is 393, the maximum number of triples (e.g., hasType) is 1,248, and the total number of triples is 7,253. For each triple, 12 prompts are generated by the templates of the associated relation. After receiving the prompts, a PLM predicts the ObjectEntity by returning a list of ranked answers in the PLM's vocabulary. All the benchmark data and experimental code are available in the public github repository (link provided in Footnote 2).

Evaluation Metric. In our benchmark dataset, a triple can be correctly filled up by one or more ObjectEntitys, each of which can contain one or more tokens, or even a SMILES string representing a molecular structure. We use top-k accuracy (acc@k), which is 1 if any of the top-k answers returned by a PLM matches an ObjectEntity, and is 0 otherwise. We use exact case-insensitive string matching to determine whether a returned answer matches an ObjectEntity. Formally, let M be the total number of triples for probing a relation, let N be the number of triples that have an ObjectEntity matched by a predicted answer in the top-k list returned by a PLM. Then, $acc@k = \frac{N}{M}$. We evaluate the performance by acc@k, where $k \in \{1, 5, 10, 50\}$.

VIII. Results and Discussion

Results. We first evaluate the overall performance of the chosen PLMs in terms of predicting MOF related triples. Figure 3 shows the average acc@1, acc@5, acc@10, and acc@50 of each PLM over all the triples. The figure clearly indicates that the domain-specific PLM, MatBERT, outperformed all the other PLMs in all the metrics.

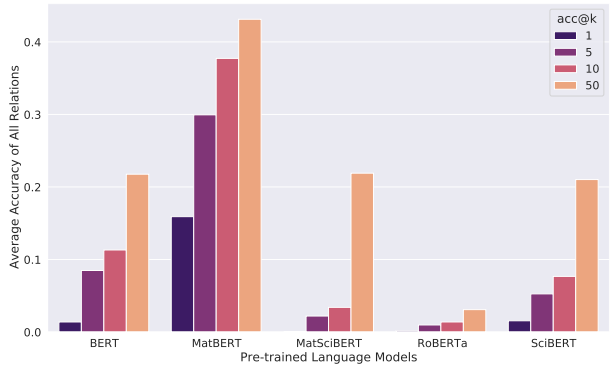

Fig. 3. For 5 individual PLMs: the average accuracy at top-k of 7 MOF-KG relations, k = 1, 5, 10, 50

Next, we evaluate the performance of the chosen PLMs on predicting MOF related triples for individual relations. Figure 4 show the acc@1, acc@5, acc@10, and acc@50 of each PLM predicting the triples of the 7 individual relations. The figure indicates that MatBERT outperformed other PLMs in 4 out 7 cases for all metrics. For the relation hasFamily, the BERT large-cased model has the best performance. For the relation hasTopology, the SciBERT has better performance on acc@5 and acc@10 than MatBERT, while MatBERT outperformed SciBERT on acc@50. Finally, no PLM could correctly predict any ObjectEntitys for the relation hasLinker. Since all the ObjectEntitys of the relation hasLinker are SMILES strings such as $[O-]C(=O)c1ccc(cc1)c1ccc(cc1)C(=O)[O-]$, all the current PLMs are limited in predicting such an object. An interesting test for future domain-specific PLMs would be to have special processing of SMILES strings or other non-natural language inclusions such as mathematical formulas.

Table IV shows the performance comparison of several published benchmarking systems for knowledge base construction from PLMs. All the systems are suffixed with 'LAMA' which stands for LAnguage Model Analysis. The original baseline LAMA [38] system probes general knowledge with manually-designed prompts. BioLAMA [45] is a benchmarking system on biological knowledge with both manually-designed (marked as `Manual_BioLAMA`) and automatically-learned prompts (marked as `Opti._BioLAMA`). MedLAMA [32] developed a benchmark from the Unified Medical Language System (UMLS) Metathesaurus. For probing PLMs, MedLAMA applied a self-supervised contrastive approach that adjusted the underlying PLMs. The last row, MOF-KG LAMA, is our benchmarking system on probing MOF-related knowledge from PLMs. Each benchmarking system probed multiple general and domain-specific PLMs. We extracted the best available acc@k results achieved by a PLM from the respective publications. The values of acc@1 show that there is still a significant gap between constructing general knowledge bases from PLMs (the baseline) vs. domain-specific ones (all the other systems).

Benchmarking System	Best acc@1	Best acc@5	Best acc@10
Baseline LAMA	**0.32**	N/A	N/A
Manual_BioLAMA	0.12	0.26	N/A
Opti._BioLAMA	0.11	0.25	N/A
MedLAMA	0.08	N/A	0.30
MOF-KG LAMA	0.16	**0.30**	**0.37**

TABLE IV
THE PERFORMANCE COMPARISON OF THE PUBLISHED BENCHMARKING LAMA SYSTEMS ON GENERAL-PURPOSE AND SPECIFIC-DOMAIN KNOWLEDGE BASE CONSTRUCTION. A BOLDFACED NUMBER IS THE BEST ACCURACY AT ITS CORRESPONDING RANK.

For each relation, we investigated the prompt that was used to probe a PLM which generated the best accuracy at the highest possible rank. Table V lists such prompt templates for all the relations. For example, for the relation hasType, the PLM MatBERT achieved the best accuracy at top-1 being probed with the prompt template *"`[SUB]` is an `[MASK]` made of metal center and organic linkers."*

Discussion. All the prompts were manually designed based on human experience and trial-and-error. Research [19], [32], [45] has indicated that prompts have significant impacts on the performance of PLMs for predicting knowledge triples. It would be more advantageous if prompts could be developed in a systematic and general way for future KBC-LM tasks. In the current study, answers are evaluated by simple exact string matching to the triples' ObjectEntitys. More sophisticated answer processing could be developed to extract the correct answers from the output space of a PLM. Researchers have

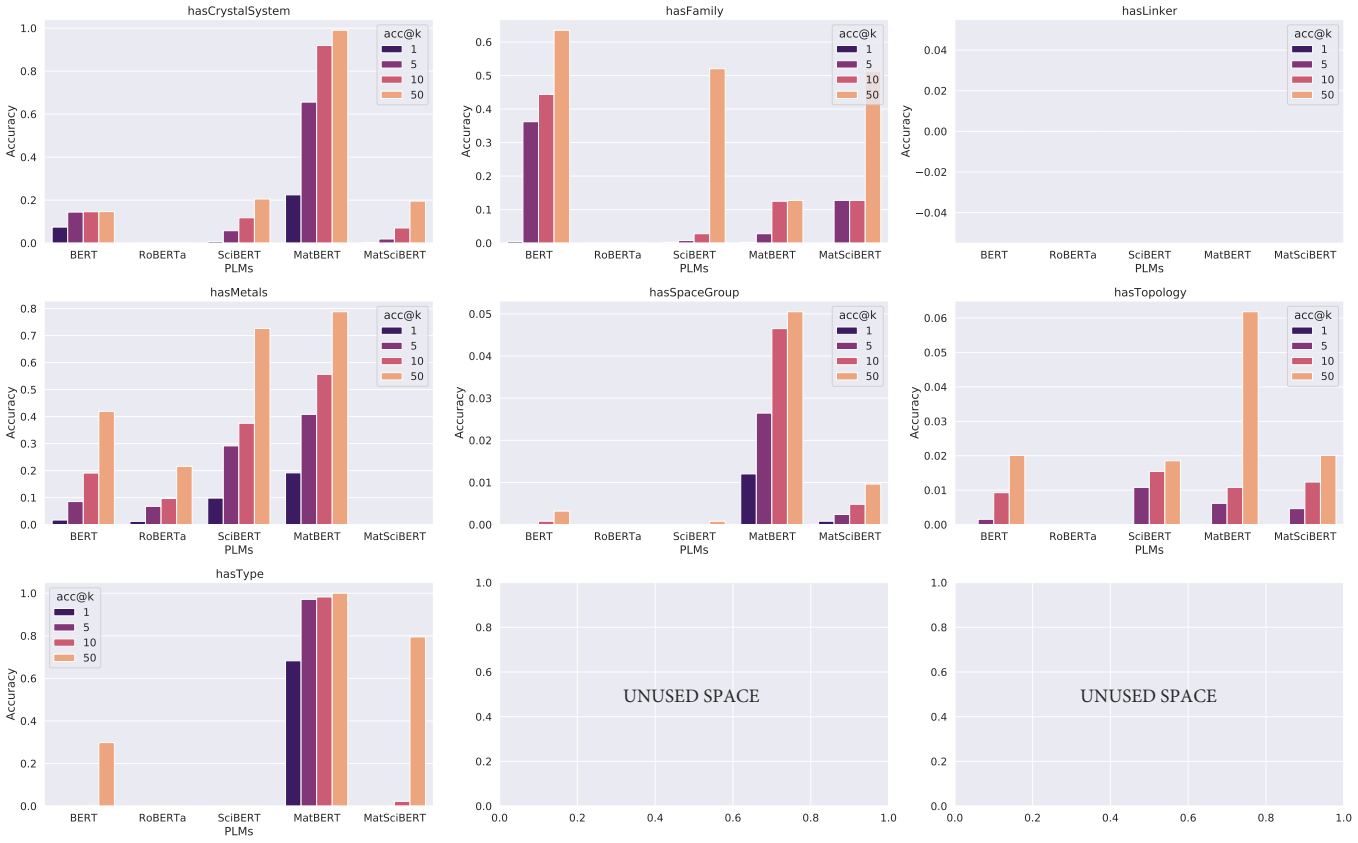

Fig. 4. For each of the 7 MOF-KG relations: the accuracy at top-k generated by each individual PLM, k = 1, 5, 10, 50; the empty plot for hasLinker indicates that no PLM could predict any results for the relation hasLinker; the last two empty plots are unused space.

No.	Relation	top-k	PLM	Prompt Template
1.	hasType	1	MatBERT	`[SUB]` is an `[MASK]` made of metal center and organic linkers.
2.	hasMetals	1	MatBERT	`[MASK]` was used as metal center in the synthesized `[SUB]` material.
3.	hasOrganicLinker	N/A	N/A	No template and PLM predicted the linker.
4.	hasFamily	1	BERT	`[SUB]` is a type of `[MASK]` MOF.
5.	hasCrystalSystem	1	MatBERT	`[SUB]` is an metal organic framework. `[SUB]`'s crystal system is `[MASK]`.
6.	hasTopology	5	SciBERT	`[SUB]` is an MOF. `[SUB]` is an type of `[MASK]` topology.
7.	hasSpaceGroup	1	MatBERT	`[SUB]` has SBUs and organic linkers. The space group of `[SUB]` is `[MASK]`.

TABLE V
THE MOST ACCURATE TEMPLATE/PLM COMBINATION FOR EACH RELATION (AT THE HIGHEST POSSIBLE RANK K)

developed manual approaches using verbalizers [14], [25], [38], [51] and automatic methods through paraphrasing [26], pruning [43], and label decomposition [12]. It is also worth noting that different PLMs had different performance on individual relations. It would be more effective to develop a learning strategy for choosing the most relevant PLM or creating an ensemble of PLMs for probing. Unexpectedly, we see that RoBERTa's performance is the worst, indicating the danger of assuming that the new and improved model is in fact always an improvement.

IX. Conclusion and Future Work

We developed a LAMA benchmark to probe pre-trained language models for constructing a knowledge graph of metal-organic frameworks (MOFs), an emerging material that has a game-changing capacity for many applications. Efficient and holistic structured knowledge integration related to MOFs would greatly assist domain experts in screening, designing, and synthesizing them. We explored a set of state-of-the-art pre-trained general-purpose and domain-specific language models. The probing results showed that a domain-specific PLM, MatBERT, consistently outperformed other general or specific PLMs for predicting MOF related triples.

The overall results, however, indicate that using the present PLMs to create domain-specific knowledge graphs is still far from being practical. This shortcoming has also been demonstrated by other LAMA benchmarking results in biological and medical domains. The study leads us to several future directions for improvements:

- Expanding the study to test more PLMs including dis-

tilled PLMs for computational efficiency and extract more knowledge graph facts such as synthesis procedures.
- Developing more capable and knowledgeable pre-trained language models for particular applications in materials science.
- Developing approaches for fine-tuning on PLMs and prompt-tuning for probing.
- Investigating automatic methods that can learn appropriate prompts by matching training triples to text corpora.
- Developing machine learning approaches that can automatically choose the most appropriate PLM for a particular type of knowledge or combining several PLMs as an ensemble.
- Developing more effective strategies to extract answers from the sets of tokens returned by a PLM or an ensemble of PLMs.

ACKNOWLEDGMENTS

This project is partially supported by the Drexel Office of Faculty Affairs' 2022 Faculty Summer Research awards #284213, and the U.S. National Science Foundation Office of Advanced Cyberinfrastructure (OAC) Grant #1940239 #2118201 and #1940307.

REFERENCES

[1] A. Aghajanyan, D. Okhonko, M. Lewis, M. Joshi, H. Xu, G. Ghosh, and L. Zettlemoyer. HTLM: Hyper-Text Pre-Training and Prompting of Language Models. *arXiv e-prints*, page arXiv:2107.06955, July 2021.

[2] D. Alivanistos, S. B. Santamaría, M. Cochez, J.-C. Kalo, E. van Krieken, and T. Thanapalasingam. Prompting as probing: Using language models for knowledge base construction. In *the Knowledge Base Construction from Pre-trained Language Models (LM-KBC) Challenge @ 21st International Semantic Web Conference*, 2022.

[3] Y. An, J. Greenberg, X. Zhao, X. Hu, S. McCLellan, A. Kalinowski, F. J. Uribe-Romo, K. Langlois, J. Furst, D. A. Gómez-Gualdrón, F. Fajardo-Rojas, and K. Ardila. Building open knowledge graph for metal-organic frameworks (mof-kg): Challenges and case studies. In *International Workshop on Knowledge Graphs & Open Knowledge Network (OKN) Co-located with the 28th ACM SIGKDD Conference*, Washington, DC, August 15, 2022, Washington, DC.

[4] R. Anderson and D. Gómez-Gualdrón. Increasing topological diversity during computational "synthesis" of porous crystals: how and why. *CrystEngComm*, 21(10):1653–1665, 2019.

[5] S. Arora. A Survey on Graph Neural Networks for Knowledge Graph Completion. *arXiv e-prints*, page arXiv:2007.12374, July 2020.

[6] H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, S. Piao, J. Gao, M. Zhou, and H.-W. Hon. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[7] S. Batten, N. Champness, X.-M. Chen, J. Garcia-Martinez, S. Kitagawa, L. Öhrström, M. O'keeffe, M. Paik Suh, and J. Reedijk. Terminology of metal–organic frameworks and coordination polymers (IUPAC Recommendations 2013). *Pure and Applied Chemistry*, 85(8):1715–1724, 7 2013.

[8] I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[9] E. Ben-David, N. Oved, and R. Reichart. PADA: Example-based Prompt Learning for on-the-fly Adaptation to Unseen Domains. *Transactions of the Association for Computational Linguistics*, 10:414–433, 04 2022.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[11] B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik, and R. Q. Snurr. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*, 19(11):6682–6697, 2019.

[12] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2778–2788, New York, NY, USA, 2022.

[13] D. K. Choe and E. Charniak. Parsing as language modeling. In *EMNLP*, pages 2331–2336, Austin, Texas, Nov. 2016.

[14] L. Cui, Y. Wu, J. Liu, S. Yang, and Y. Zhang. Template-based named entity recognition using BART. In *ACL-IJCNLP*, pages 1835–1845, Online, Aug. 2021.

[15] J. Davison, J. Feldman, and A. Rush. Commonsense knowledge mining from pretrained models. In *EMNLP-IJCNLP*, pages 1173–1178, Hong Kong, China, Nov. 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[17] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022.

[18] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon. *Unified Language Model Pre-Training for Natural Language Understanding and Generation*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[19] X. Fang, A. Kalinowski, H. Zhao, Z. You, Y. Zhang, and Y. An. Prompt design and answer processing for knowledge base construction from pre-trained language models (lm-kbc). In *the Knowledge Base Construction from Pre-trained Language Models (LM-KBC) Challenge @ 21st International Semantic Web Conference*, 2022.

[20] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. In *ACL*, pages 3816–3830, Online, Aug. 2021.

[21] T. Gupta, M. Zaki, N. Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 5 2022.

[22] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun. Ptr: Prompt tuning with rules for text classification. *arXiv preprint arXiv:2105.11259*, 2021.

[23] A. Haviv, J. Berant, and A. Globerson. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online, Apr. 2021.

[24] T. He, K. Cho, and J. R. Glass. An empirical study on few-shot knowledge probing for pretrained language models. *CoRR*, abs/2109.02772, 2021.

[25] Z. Jiang, A. Anastasopoulos, J. Araki, H. Ding, and G. Neubig. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *EMNLP*, pages 5943–5959, Online, Nov. 2020.

[26] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 07 2020.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880, Online, July 2020.

[28] C. Li, F. Gao, J. Bu, L. Xu, X. Chen, Y. Gu, Z. Shao, Q. Zheng, N. Zhang, Y. Wang, and Z. Yu. SentiPrompt: Sentiment Knowledge Enhanced Prompt-Tuning for Aspect-Based Sentiment Analysis. *arXiv e-prints*, page arXiv:2109.08306, Sept. 2021.

[29] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, pages 4582–4597, Online, Aug. 2021.

[30] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang. GPT understands, too. *CoRR*, abs/2103.10385, 2021.

[31] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692, July 2019.

[32] Z. Meng, F. Liu, E. Shareghi, Y. Su, C. Collins, and N. Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *60th ACL*, 2022.

[33] P. Moghadam, A. Li, S. Wiggin, A. Tao, A. Maloney, P. Wood, S. Ward, and D. Fairen-Jimenez. Development of a cambridge structural database subset: A collection of metal–organic frameworks for past, present, and future. *Chemistry of Materials*, 29(7):2618–2625, 3 2017.

[34] P. Z. Moghadam, A. Li, X.-W. Liu, R. Bueno-Perez, S.-D. Wang, S. B. Wiggin, P. A. Wood, and D. Fairen-Jimenez. Targeted classification of metal–organic frameworks in the cambridge structural database (csd). *Chem. Sci.*, 11:8373–8387, 2020.

[35] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, Apr. 2019.

[36] X. Ouyang, S. Wang, C. Pang, Y. Sun, H. Tian, H. Wu, and H. Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *EMNLP*, pages 27–38, Online and Punta Cana, Dominican Republic, Nov. 2021.

[37] M. O'keeffe, M. Peskov, S. Ramsden, and O. Yaghi. The Reticular Chemistry Structure Resource (RCSR) Database of, and Symbols for, Crystal Nets. *Accounts of Chemical Research*, 41(12):1782–1789, 12 2008.

[38] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller. Language models as knowledge bases? In *EMNLP-IJCNLP*, pages 2463–2473, Hong Kong, China, Nov. 2019.

[39] N. Poerner, U. Waltinger, and H. Schütze. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. *arXiv e-prints*, page arXiv:1911.03681, Nov. 2019.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[41] T. Schick and H. Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–269, Online, Apr. 2021.

[42] T. Schick and H. Schütze. It's not just size that matters: Small language models are also few-shot learners. In *NAACL*, pages 2339–2352, Online, June 2021.

[43] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of EMNLP*, pages 4222–4235, Online, Nov. 2020.

[44] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936, 2019.

[45] M. Sung, J. Lee, S. Yi, M. Jeon, S. Kim, and J. Kang. Can language models be biomedical knowledge bases. In *EMNLP*, 2021.

[46] A. Trewartha, N. Walker, H. Huo, S. Lee, K. Cruse, J. Dagdelen, A. Dunn, K. A. Persson, G. Ceder, and A. Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4):100488, 2022.

[47] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for attacking and analyzing NLP. In *EMNLP-IJCNLP*, pages 2153–2162, Hong Kong, China, Nov. 2019.

[48] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun. Pre-trained language models and their applications. *Engineering*, 2022.

[49] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702, 2019. PMID: 31361962.

[50] O. Yaghi. Reticular chemistry in all dimensions. *ACS Central Science*, 5(8):1295–1300, 8 2019.

[51] W. Yin, J. Hay, and D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *EMNLP-IJCNLP*, pages 3914–3923, Hong Kong, China, Nov. 2019.

[52] W. Yuan, G. Neubig, and P. Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *NeurIPS*, volume 34, pages 27263–27277, 2021.

[53] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced language representation with informative entities. In *ACL*, pages 1441–1451, Florence, Italy, July 2019.

[54] X. Zhao, J. Greenberg, S. McClellan, Y.-J. Hu, S. Lopez, S. K. Saikin, X. Hu, and Y. An. Knowledge graph-empowered materials discovery. In *1st Workshop on Knowledge Graph and Big Data collocated with 2021 IEEE International Conference on Big Data (Big Data)*, 2021.

[55] Z. Zhong, D. Friedman, and D. Chen. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL*, 2021.