Building Open Knowledge Graph for Metal-Organic Frameworks (MOF-KG): Challenges and Case Studies

Yuan An, Jane Greenberg, Xintong Zhao, Xiaohua Hu, Scott McCLellan, Alex Kalinowski
College of Computing and Informatics
Drexel University
Philadelphia, PA, USA
Contact:{ya45,jg3243,xh29}@drexel.edu

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

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

***Abstract*—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 Cambridge Structural Database (CSD) reports 10,636 synthesized MOF crystals which in addition contains ca. 114,373 MOF-like structures. The sheer number of synthesized (plus potentially synthesizable) MOF structures requires researchers pursue computational techniques to screen and isolate MOF candidates. In this demo paper, we describe our effort on leveraging knowledge graph methods to facilitate MOF prediction, discovery, and synthesis. We present challenges and case studies about (1) construction of a MOF knowledge graph (MOF-KG) from structured and unstructured sources and (2) leveraging the MOF-KG for discovery of new or missing knowledge.**

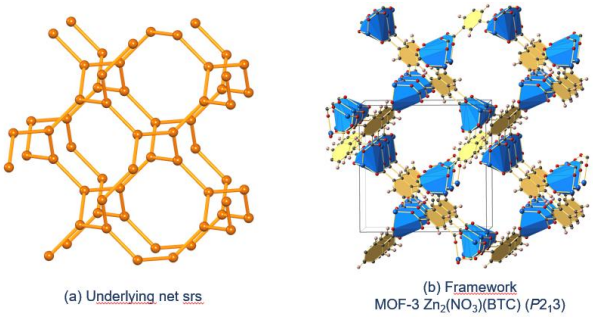


Fig. 1: The Underlying Net and Framework of MOF-3

I. INTRODUCTION

Metal-Organic Frameworks (MOFs) are materials that possess modular, porous crystal structures that can be (conceptually) modified by "swapping" constituent building blocks. MOF building blocks correspond to metal-based clusters and organic linkers, which are interconnected in patterns described by an underlying net (Figure 1). MOFs have great potential to revolutionize applications such as gas storage, molecular separations, sensing, catalysis, and drug delivery, primarily due to their usually high surface area and exceptionally tunable properties [1]. But the combinatorics of building blocks means that chemists have access to a (not fully explored) "material design space" of trillions of structures. To date, there are 10,636 synthesized MOF crystals reported in the Cambridge Structural Database (CSD) [2] which in addition contains ca. 114,373 MOF-like structures. The sheer number has made the identification of optimal MOFs (and subsequent) synthesis for a given application a very challenging task. Thus, considerable efforts have been put into developing effective computational techniques to screen and isolate candidate MOF structures for the application of choice.

Previous efforts include the creation of large MOF databases and development of high throughput automated workflows such as molecular simulation and machine learning for MOF property prediction. MOF databases contains both synthesized and "hypothesized" MOF structures [3]. When a hypothesized MOF structure is identified as promising, it is unclear if

and/or how such structure can be realized synthetically. This is partly due to the crystallization process leading to MOF formation not being fully understood. Worse, a large amount of MOF synthesis data are not readily available for computation but scattered in scientific literature. The separation between the crystals' structure information and their synthesis data exacerbates the difficulty of screening MOFs.

We have undertaken a project aiming to leverage advanced knowledge graph methods to facilitate MOF prediction, discovery, and synthesis. In this demo paper, we describe the challenges for building such an open knowledge graph for MOFs from structured and unstructured data. We present case studies on addressing several challenges.

Knowledge graph (KG) represents the knowledge in a domain using a graphical structure consisting of (typed-) vertices and (typed-) links. Despite the growing number of materials science related databases [4], knowledge graphs built for materials science domains are still rare [5], [6]. There is no applicable knowledge graph for MOFs or Reticular Chemistry in general. The ***first challenge*** we face in building such a knowledge graph is to choose the underlying graph architecture, for example, whether using RDF triple store or labeled property graph model. Materials science data typically describe numeric electronic, chemical, and physical properties. These properties may connect through different relationships for guiding the design of multifunctional materials. Additional properties can be extracted from the literature. Considering the need of capturing the properties, we choose the labeled property graph model for the underlying architecture and implement it in the Neo4j platform.

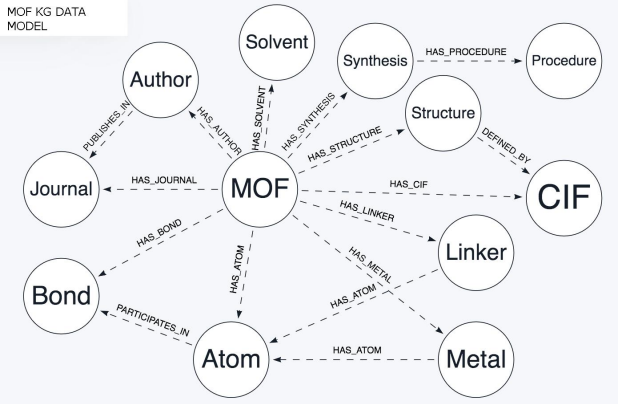


Fig. 2: Partial Backbone of the Data Model for the MOF-KG.

II. DEFINING A DATA MODEL FOR THE MOF KNOWLEDGE GRAPH (MOF-KG)

The knowledge graph for MOFs (MOF-KG) should facilitate researchers in reticular chemistry to practice their distinguished activities. The past decades have seen an explosive increase in synthesis and characterization of MOFs. A systematic workflow pattern in the field has emerged [7]. The pattern is an iterative cycle consisting of 3 refinement steps: *Synthesis, Activation*, and *Analysis*. The workflow typically starts with a *Synthesis* step where the researchers will screen and identify various crystalline structures by considering many parameters including starting compounds, molar ratios, temperature, concentration, reaction additives, modulators, solvents, and reaction time. The researchers can apply computational techniques such as high-throughput synthesis or screening to aid their experimental process [8].

Next at *Activation*, the researchers will assess the permanent porosity and architectural stability of the crystal by removing all guest molecules (including solvent) from the pores of the framework without causing collapse of its structure. The third step is *Analysis* where the researchers will identify and characterize the physical and chemical properties of the crystal. Here, the researchers may resort to analytical, computational, and machine learning tools to study and characterize porosity [9], gas uptake [10], stability [11], and other important properties [12]. The workflow will cycle through these 3 steps iteratively until successful results are obtained.

Due to the rapid development in reticular chemistry, there is no a general agreed system of nomenclature for describing MOFs and related activities. A couple of initiatives have worked on standardizing terminologies [13], [14], although the diversity in the focus and the scientific inquiry has led to a variety of terminological usages for this class of compounds. The ***second challenge*** in building the MOF-KG is to define a data model to conceptualize MOF and its related activities corresponding to the workflow pattern.

To address this challenge, we define a data model for the MOF-KG with 4 major areas: *synthesis, structure, atomic composition*, and *publication*. Figure 2 illustrates a partial backbone of the data model with main concepts and relationships. Each concept and relationship has its own set of properties. The data model is continuously being refined and extended with newly identified concepts and additional relationships.

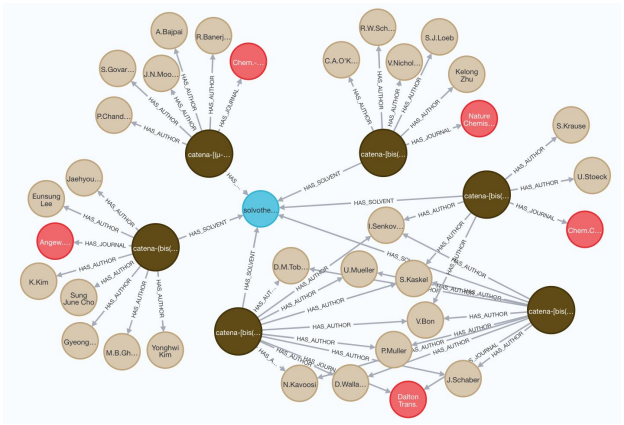


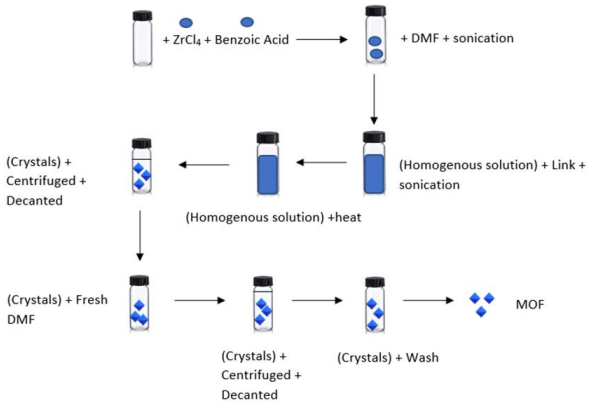
Fig. 3: Excerpt of the MOF-KG Showing the Connections between Journals, Authors, the Same Solvent, and MOFs.

III. BUILDING THE MOF-KG FROM THE CSD MOF COLLECTION

To instantiate the data model with instances for building the KG, we extract MOF related entities, relationships, and properties from both structured and unstructured sources. For

Synthesis of PIZOF-*n*: The syntheses of the PIZOFs were performed in 100 mL Teflon-capped glass flasks. For a typical synthesis, $ZrCl_4$ (0.080 g, 0.343 mmol) and benzoic acid (1.256 g, 10.287 mmol) were dissolved in DMF (20 mL) by using ultrasound to give a clear solution. One of the di-acids $HO_2C[PE\text{-}P(R^1,R^2)\text{-}EP]CO_2H$ **1b–8b** (0.343 mmol) was added and dissolved by the application of ultrasound. The tightly capped flask was kept in an oven at 120 °C under static conditions for 24 h. The suspension was cooled to room temperature and the precipitate was isolated by centrifugation. The solid was suspended in DMF (10 mL). After standing at room temperature for 1–4 h, the suspension was centrifuged and the solvent was decanted off. The obtained particles were washed with ethanol (10 mL) in the same way as described for DMF. Finally, the solid was dried under reduced pressure.

(a) A published paragraph describing a synthesis procedure



(b) The extracted synthesis procedure steps from the paragraph on left

Fig. 4: Extracting synthesis procedure steps from synthesis paragraph.

structured sources, we currently focus on 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. An important feature of the CSD is that it contains successful structures (crystals) that have been realized in experiment, with crystal structure experimentally measured and solved using diffraction techniques (X-rays, neutrons, electrons). The database contains data such as crystal symmetry, atom positions, occupancy, and displacements in the form of a CIF (crystallographic information file) which is a highly standard format for storing crystallographic structural data and metadata. The ***third challenge*** for us is to map the structured information in CSD to the MOF-KG described by the data model. To address the challenge, we develop a schema mapping tool to (semi-)automatically extract, transform, and load data from CSD to the MOF-KG.

Once the MOF-KG is populated with the instances extracted from the CSD database, we can issue queries against the knowledge graph to explore connections among the MOF related entities. Figure 3 shows an excerpt from the resultant graph after issuing the following query: *"retrieve the authors and publication journals of the MOFs that have the same solvent."*

IV. AUGMENTING THE MOF-KG WITH SCHOLARLY ARTICLES

While the curated databases such as CSD contain significant amounts of information about crystal structures, scholarly articles that are rapidly growing in quantities contain rich knowledge resources of the synthesis procedures. However, computers cannot recognize the sequence of synthesis actions reported in plain text (Figure 4a) [15]. Hence, the ***fourth challenge*** is to extract synthesis conditions from unstructured text to augment the MOF-KG (Figure 4b). There is seldom a standardized way for reporting synthesis conditions. Every researcher presents synthesis in different manners, e.g., in main article, implicit or explicit, in appendix or supplementary information file. Also, these conditions are often reported incomplete, as many researchers have pre-assumptions of their readership. This incomplete/assumed information may hinder immediate reproducibility of the synthesis.

Some effort has been put into the use of natural language processing (NLP) techniques [16], [17], [18], [19], [20], [21] for synthesis extraction. Existing techniques include rule-based (e.g., regular expression, pre-defined rules) and deep learning/machine learning based approaches [20], [21]. As a case study, we apply a recently developed rule-based NLP approach [19] to extract the synthesis information from 114 articles that match 114 MOFs in the CSD Collection. There are 46 synthesis routes were recognized and they turned out to be accurate. However, only three solvent records were extracted. We manually examined ten synthesis procedures reported in articles. The results indicate that solvents are indeed reported in text but not recognized by the extraction algorithm. One possible reason is that solvent information is described in various contexts and different ways (e.g., water, N,N-dimethyl formamide, DMF).

Rule-based approaches are convenient to implement but brittle to the text context. They suffer from the low recall problem. Deep learning and general machine learning based approaches show better performance than rule-based approaches, but they suffer from the lack of annotated training corpus. Currently, we are developing weakly-supervised information extraction algorithms to address this problem.

V. USING THE MOF-KG FOR DISCOVERING NEW AND MISSING KNOWLEDGE

Knowledge graphs often suffer from incompleteness. For example, even the state-of-the-art NLP techniques cannot extract all available information from text. A flurry of research has been conducted on knowledge graph completion by predicting missing links [22]. A notable approach is to learn knowledge graph representations, that is, low-dimensional embedding vectors for downstream classification and prediction [23]. The MOF-KG is inevitably incomplete due to either

incomplete information in the CSD database or incomplete extraction from text. We aim to develop MOF-KG related link prediction techniques for discovering new and missing information.

As a case study, we choose to predict the `'HAS_SOLVENT'` links in the MOF-KG. Solvent information is an important variable for MOF synthesis and 97% of the solvent information is missing in the CSD MOF collection. We extract the publication, author, atom, and bond information for the 268 MOFs that have solvent information in the CSD collection. We apply various knowledge graph embedding models including *TransE, ConvE, ComplEx, DistMult* and *SimplE* on the data set. We randomly sample 20% as the test data and predict the `'HAS_SOLVENT'` relation for them. The results are reported in Table I. We use the following rank-based evaluation metrics:

- MRR (Inverse Harmonic Mean Rank): higher is better, range $[0, 1]$.
- Hits@K (with K as one of $\{1, 5, 10\}$): higher is better, range $[0, 1]$.
- AMRI (Adjusted Arithmetic Mean Rank Index): higher is better, range $[-1, 1]$. AMRI=0 means the model is not better than random scoring.

KGE Model	MRR	AMRI	Hits@10	Hits@5	Hits@1
TransE	0.15	0.97	0.38	0.25	0.04
ConvE	0.12	0.94	0.21	0.18	0.07
ComplEx	0	-0.13	0	0	0
DistMult	**0.25**	**0.99**	**0.5**	**0.42**	**0.14**
SimplE	0	0.11	0	0	0

TABLE I: Apply Various Knowledge Graph Embedding (KGE) Models for Predicting Missing Solvent information for MOFs

The results show that the DistMult KG embedding model achieved the best result in terms of all the metrics. The AMRI is close to 1 and Hits@10 is 0.5. With the simple information of publication and atomic element, a KG embedding can successfully predict important missing information. Encouraged by these results, we plan to develop and apply more sophisticated link prediction approaches to the MOF-KG.

VI. CONCLUSION AND NEXT STEPS

The challenges presented here show that building a knowledge graph for MOFs requires a strong synergy among materials scientists, chemists, informaticians, and data and computer scientists. The case studies demonstrate the potentials of knowledge graphs for discovering information that is missing in the original distributed and heterogeneous sources. This project is a part of the effort undertaken by the NSF Institute for Data Driven Dynamical Design (ID4). The next steps include enriching the MOF-KG and building user-friendly natural language and chatbot query interfaces for domain scientists to conduct further knowledge graph-empowered materials discovery.

ACKNOWLEDGMENT

The research reported on in this paper is supported, in part, by the U.S. National Science Foundation, Office of Advanced Cyberinfrastructure (OAC): Grant: 1940239 and 1940307.

REFERENCES

[1] Yaghi OM. Reticular Chemistry in All Dimensions. ACS Cent Sci. 2019;5(8):1295-1300. doi:10.1021/acscentsci.9b00750

[2] Moghadam, Peyman, et al. Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. Chem. Mater. 2017, 29, 7, 2618–2625.

[3] Anderson, R. and Gomez-Gualdron D.A., Increasing topological diversity during computational "synthesis" of porous crystals: how and why, CrystEngComm (2019), 21, 1653

[4] Zhao, Xintong, et al. "Knowledge Graph-Empowered Materials Discovery." 2021 IEEE International Conference on Big Data (Big Data). IEEE, (2021).

[5] Mrdjenovich, David, et al. "propnet: A knowledge graph for materials science." Matter 2.2 (2020): 464-480.

[6] McCusker, James P., et al. "Nanomine: A knowledge graph for nanocomposite materials science." International Semantic Web Conference. Springer, Cham, 2020.

[7] Gropp C, Canossa S, Wuttke S, et al. Standard Practices of Reticular Chemistry. ACS Cent Sci. 2020;6(8):1255-1273.

[8] Davariashtiyani, A., Kadkhodaie, Z. & Kadkhodaei, S. Predicting synthesizability of crystalline materials via deep learning. Commun Mater 2, 115 (2021).

[9] Peyman Z. Moghadam, et al. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. Matter, (2019).

[10] Tayfuroglu, Omer et al. "In silico investigation into H2 uptake in MOFs: combined text/data mining and structural calculations." Langmuir : the ACS journal of surfaces and colloids (2019).

[11] Nandy, A., Terrones, G., Arunachalam, N. et al. MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. Sci Data 9, 74 (2022).

[12] Rosen, Andrew, et al. Machine Learning the Quantum-Chemical Properties of Metal–Organic Frameworks for Accelerated Materials Discovery with a New Electronic Structure Database. Matter. May 2021.

[13] O'Keeffe M, Peskov MA, Ramsden SJ, Yaghi OM. The Reticular Chemistry Structure Resource (RCSR) database of, and symbols for, crystal nets. Acc Chem Res. 2008 Dec;41(12):1782-9.

[14] Batten, Stuart R., et al. Terminology of metal–organic frameworks and coordination polymers (IUPAC Recommendations 2013). Pure and Applied Chemistry, vol. 85, no. 8, 2013, pp. 1715-1724.

[15] Schaate, A., et al. Porous Interpenetrated Zirconium–Organic Frameworks (PIZOFs): A Chemically Versatile Family of Metal–Organic Frameworks. Chem. Eur. J. 2011, 17, 9320.

[16] Kononova, O., Huo, H., He, T. et al. Text-mined dataset of inorganic materials synthesis recipes. Sci Data 6, 203 (2019).

[17] He, Tanjin et al. Similarity of Precursors in Solid-State Synthesis as Text-Mined from Scientific Literature. Chemistry of Materials 32 (2020): 7861-7873.

[18] Huo, H., Rong, Z., Kononova, O. et al. Semi-supervised machine-learning classification of materials synthesis procedures. npj Comput Mater 5, 62 (2019).

[19] Kristian Gubsch, et al. DigiMOF: A Database of MOF Synthesis Information Generated via Text Mining. Chemarxiv. April (2022).

[20] Mysore, Sheshera et al. Automatically Extracting Action Graphs from Materials Science Synthesis Procedures. arXiv preprint arXiv:1711.06872, (2017).

[21] Huang, Shu and Jacqueline M Cole. A Database of Battery Materials Auto-Generated Using Chemdataextractor. Scientific data, (2020).

[22] Andrea Rossi, et al. Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. ACM Trans. Knowl. Discov. (2021)

[23] Galkin, M., Berrendorf, M., and Tapley Hoyt, C., "An Open Challenge for Inductive Link Prediction on Knowledge Graphs", arXiv e-prints, (2022).