MOFology: Ontologically-Grounded MOF Knowledge Graph and Prediction Framework

Matthew Hart [1] **Philippe Schwaller** [1]

[1]*Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland*. Correspondence to: Matthew Hart matthew.hart@epfl.ch.

1. Introduction

Metal-Organic Frameworks (MOFs) have emerged a chemically versatile material platform, with demonstrated utility in gas separation, catalysis, and many other areas.[1] Their modular construction consisting of metal nodes connected by organic linkers gives rise to a quasi-infinite combinatorial design space that enable their chemical tuneability.[1] These factors have motivated substantial experimental and computational study, producing large datasets and growing volumes of literature.[2] This situation presents both a challenge and an opportunity to materials chemists; information overload on one hand, and the possibility of leveraging this information to design useful materials on the other.[3]

Despite the volume of available MOF data, most machine learning studies in the field remain focused on a few databases and at most a handful of numerical properties, with both generative and predictive models[4, 5] employed to design and screen MOFs.[6] While these studies have demonstrated individual success, the models they produce have little ability to draw on the broader relational context that connects a framework's topology, linker chemistry, synthesis route, and application performance. This limitation is the result of MOF data being fragmented, with data on each modality spread across several databases and journal articles. As a result, the collective knowledge embedded across MOF databases is greater than what any individual model can currently access.

Knowledge graphs (KGs) and ontologies[7] offer a natural solution to the fragmentation of MOF data. These frameworks have been widely used across various fields to integrate databases by representing entities and the relationships between them as structured graphs.[8] This system enables semantic searching across the data contained in multiple databases, and oftentimes reveals connections between entities that were not explicitly stated in the original sources.[7, 9] However, the value of a KG depends greatly on how data is ingested and represented. Simply mapping database tables into nodes and edges produces a shallow graph that mirrors the limitations of its sources.[10] Ontologies address this by providing a formal layer acting as a shared vocabulary of entity/relationships types and taxonomic constraints. Existing literature applying knowledge graph methods to MOFs[11, 12] remain limited in scale and lack ontological integration and downstream predictive capability, creating the clear need for MOF KGs that combine the reliability of ontologies with the predictive power of machine learning.

Beyond data integration, a KG allows for the ability to explore both structure-based and semantically-based latent spaces simultaneously. Knowledge graph embeddings[13] create high dimensional representations of the the entities that exist within a KG, and allow for the integration of various machine learning methods. By combining these embeddings with structural chemical features into a shared latent space, predictive models are able to improve performance in both domains. Here, we present the largest ontologically grounded MOF knowledge graph to date, together with a prediction framework built on its joint semantic-chemical embedding. As a test case, we present the first incorporation of amine-functionalized MOFs into a KG-based predictive setting, targeting the effect of amine functionalization on binding energies for species relevant to direct air capture (DAC) technologies.

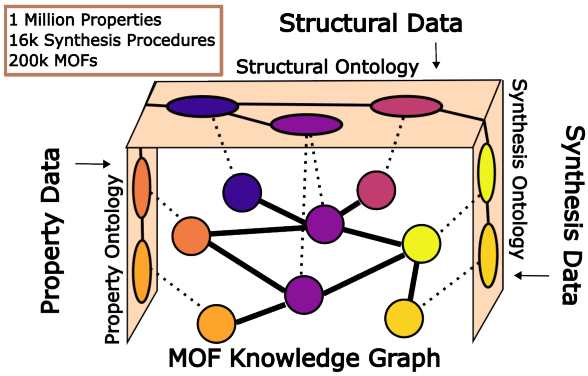

Fig. 1: Visual representation of the constructed MOF KG. The outer box represents the MOF ontology that was divided into sections specializing in synthesis structure, and properties. Colors represent entity types and dashed lines represent "is a" relationships connecting KG instances to the ontology.

2. Methodology

The KG was constructed through a pipeline of structured and unstructured data extraction, ontological alignment, and instantiation. A general schema of MOF chemistry entity linking was constructed based on a combination of the MOFKG[12] and MOFChemUnity[8] data models with the necessary modifications to include data provenance, functionalization, binding properties, and synthesis procedures. An Elementary Multiperspective Material Ontology (EMMO)[14]-based ontology was then

constructed by integrating the custom schema into the EMMO name space and adding modifications to the ontology's previously existing MOF subsection. MOF data was taken from the publicly available versions of OpenDAC25[15], MOFChemUnity[11], QMOF[16], CSD[17], DigiMOF[18], SynMoF[5], and MOF-FreeEnergy[5] representing structure, functionalization, binding energies, physical properties, structural properties, synthesis information, stability information, linker information, and several other MOF data modalities. These were then linked using the data models provided by the custom ontology and instantiated in a neo4j[19] database. The final knowledge graph had 1 million properties, 200,000 MOFs, 16,000 synthesis procedures, and 15,000 abstracts contained. As a final step, the MOASAEC algorithm[20] was used to verify the metal oxidation states of MOFS in the KG. Figure 1 contains a depiction of how elements of the knowledge graph are connected to the ontology.

For the predictive pipeline, the KG was embedded into a continuous latent space using two approaches to compare the effect of encoding strategy on downstream performance. Node2Vec[21] was used as a topology-only baseline, generating embeddings based solely on graph structure through biased random walks. These were compared against a CompGCN-based[22] graph attention architecture, which jointly encodes topological structure, node features, and edge-type semantics. Chemical descriptors were concatenated with the resulting embeddings to form a fused semantic-chemical representation. Both embedding strategies were evaluated on four downstream tasks: link prediction, data imputation, property prediction, and concept vector analysis[23].

3. Results

The two embedding strategies exhibit complementary strengths across the evaluated tasks. Node2Vec, operating on graph topology alone, performs competitively on simpler physical properties where structural connectivity is the dominant predictor. However, CompGCN consistently outperforms Node2Vec on KG-specific tasks such as link prediction (AUC 0.89) and data imputation as well as on more complex property prediction targets where semantic and relational context contribute meaningfully beyond topology. Across property prediction tasks, R^2 values range from 0.5 to 0.95 depending on the target property and embedding strategy, with the largest performance gaps between the two approaches appearing for properties that depend on multiple relational factors such as synthesis conditions and functionalization. Concept vector analysis in the CompGCN embedding space reveals interpretable structure in the latent representation. Specifically, the direction corresponding to amine functionalization in the latent space aligns with a corresponding shift in predicted CO_2 binding energy, demonstrating that the embedding captures both static relationships and the effect of chemical transformations on downstream properties. This result suggests that the fused semantic-chemical space encodes actionable design information and the latent representation of functionalization carries predictive meaning for properties directly relevant to DAC applications. These concept vectors form the basis for ongoing experimental validation and point toward a generative framework in which traversal of the latent space corresponds to physically meaningful MOF design. Figure 2 shows the process of how KG embedding reveals the mentioned concept vectors and the predicted vs actual plot of amine functionalization on MOF CO_2 binding energies.

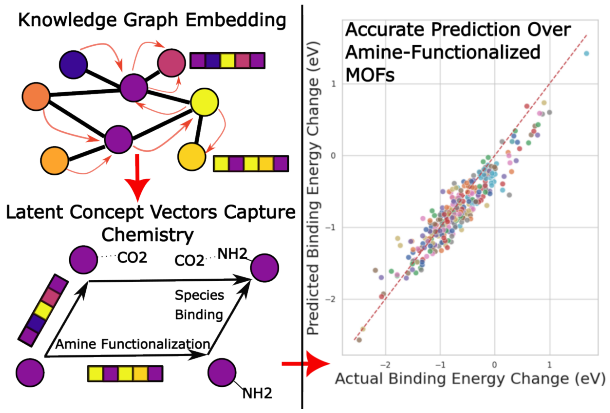

Fig. 2: Knowledge graph embedding reveals "concept vectors" where directions in the latent space correspond to real chemical changes in one direction, and species binding in another. These vectors are used to accurately predict the effect of amine functionalization on CO_2 binding energy.

4. Discussion

The knowledge graph presented here represents the largest ontologically-grounded MOF KG to date, integrating several major databases through a principled EMMO-based ontological framework that ensures entities and relationships carry disambiguated meaning. The resulting resource is both human-searchable and machine-readable, enabling researchers to query across structure, synthesis, properties, and application context in a single unified system. From a predictive standpoint, the incorporation of amine-functionalized MOFs into a KG-based prediction setting is, to our knowledge, the first such example. The emergence of functionalization as an interpretable direction in the latent space suggests that this approach captures design-relevant chemical abstractions, with direct applicability to DAC-relevant MOF screening where understanding the effect of functionalization on binding energies for CO_2, H_2O, and N_2 is central to materials design. Looking forward, the concept vectors identified here will guide experimental validation and serve as the basis for generative models operating over the joint semantic-chemical space.

Acknowledgments

MH acknowledges funding from the EPFL large-scale Solutions4Sustainability demonstrator Project (SusEcoCCUS).

References

[1] Ralph Freund, Orysia Zaremba, Giel Arnauts, Rob Ameloot, Grigorii Skorupskii, Mircea Dincă, Anastasiya Bavykina, Jorge Gascon, Aleksander Ejsmont, Joanna Goscianska, Markus Kalmutzki, Ulrich Lächelt, Evelyn Ploetz, Christian S. Diercks, and Stefan Wuttke. The current status of mof and cof applications. *Angewandte Chemie International Edition*, 60(45):23975–24001, 2021.

[2] Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, et al. Crest—a program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics*, 160(11), 2024.

[3] Jake Burner, Jun Luo, Andrew White, Adam Mirmiran, Ohmin Kwon, Peter G. Boyd, Stephen Maley, Marco Gibaldi, Scott Simrod, Victoria Ogden, and Tom K. Woo. Arc–mof: A diverse database of metal-organic frameworks with dft-derived partial atomic charges and descriptors for machine learning. *Chemistry of Materials*, 35(3):900–916, 2023.

[4] Junwu Chen, Jeff Guo, and Philippe Schwaller. Matinvent: Reinforcement learning for 3d crystal diffusion generation. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

[5] Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. Mof synthesis prediction enabled by automatic data mining and machine learning. *Angewandte Chemie International Edition*, 61(19):e202200242, 2022.

[6] Yanjing He, Zhi Fang, Wenjuan Xue, Tongan Yan, Shitong Zhang, Zhengqing Zhang, Hongliang Huang, and Chongli Zhong. Discovery of metal-organic frameworks for efficient nf3/n2 separation by integrating high-throughput computational screening, machine learning, and experimental validation. *Separation and Purification Technology*, 364:132481, 2025.

[7] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial intelligence review*, 56(11):13071–13102, 2023.

[8] Xuefeng Bai, Song He, Yi Li, Yabo Xie, Xin Zhang, Wenli Du, and Jian-Rong Li. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj Computational Materials*, 11(1):51, 2025.

[9] Raghupathi Thota, Sri Meghana Potluri, Ahmed Hassan Saker Alzaidy, P Bhuvaneshwari, et al. Knowledge graph construction-based semantic web application for ontology development. In *2025 International Conference on Intelligent Computing and Knowledge Extraction (ICICKE)*, pages 1–6. IEEE, 2025.

[10] Alexandru Oarga, Matthew Hart, Andres M Bran, Magdalena Lederbauer, and Philippe Schwaller. Scientific knowledge graph and ontology generation using open large language models. *Digital Discovery*, 2026.

[11] Thomas Michael Pruyn, Amro Aswad, Sartaaj Takrim Khan, Ju Huang, Robert Black, and Seyed Mohamad Moosavi. Mof-chemunity: Literature-informed large language models for metal–organic framework research. *Journal of the American Chemical Society*, 147(47):43474–43486, 2025.

[12] Yuan An, Jane Greenberg, Xintong Zhao, Xiaohua Hu, Scott McCLellan, Alex Kalinowski, Fernando J Uribe-Romo, Kyle Langlois, Jacob Furst, Diego A Gómez-Gualdrón, et al. Building open knowledge graph for metal-organic frameworks (mof-kg): Challenges and case studies. *arXiv preprint arXiv:2207.04502*, 2022.

[13] Jiahang Cao, Jinyuan Fang, Zaiqiao Meng, and Shangsong Liang. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Computing Surveys*, 56(6):1–42, 2024.

[14] Anne De Baas, Pierluigi Del Nostro, Jesper Friis, Emanuele Ghedini, Gerhard Goldbeck, Ilaria Maria Paponetti, Andrea Pozzi, Arkopaul Sarkar, Lan Yang, Francesco Antonio Zaccarini, et al. Review and alignment of domain-level ontologies for materials science. *IEEE Access*, 11:120372–120401, 2023.

[15] Anuroop Sriram, Logan M Brabson, Xiaohan Yu, Sihoon Choi, Kareem Abdelmaqsoud, Elias Moubarak, Pim de Haan, Sindy Löwe, Johann Brehmer, John R Kitchin, et al. The open dac 2025 dataset for sorbent discovery in direct air capture. *arXiv preprint arXiv:2508.03162*, 2025.

[16] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

[17] Frank H Allen and Robin Taylor. Research applications of the cambridge structural database (csd). *Chemical Society Reviews*, 33(8):463–475, 2004.

[18] Lawson T Glasby, Kristian Gubsch, Rosalee Bence, Rama Oktavian, Kesler Isoko, Seyed Mohamad Moosavi, Joan L Cordiner, Jason C Cole, and Peyman Z Moghadam. Digimof: a database of metal–organic framework synthesis information generated via text mining. *Chemistry of Materials*, 35(11):4510–4524, 2023.

[19] Justin J Miller. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324, pages 141–147. AISEL, 2013.

[20] Andrew J White, Marco Gibaldi, Jake Burner, R Alex Mayo, and Tom K Woo. High structural error rates in "computation-ready" mof databases discovered by checking metal oxidation states. *Journal of the American Chemical Society*, 147(21):17579–17583, 2025.

[21] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[22] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1911.03082*, 2019.

[23] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.