M-MOFormer: Multi-Modal Transformer Framework for Metal-Organic Framework Property Prediction

Boya Min[1], **Luoxiao Yang**[2*], **Zijun Zhang**[3]
[1]Xi'an Jiaotong University, Xi'an, China
[2]Xi'an University of Technology, Xi'an, China
[3]City university of Hong Kong, Hong Kong SAR
minboya@stu.xjtu.edu.cn, luoxiyang2-c@my.cityu.edu.hk, zijzhang@cityu.edu.hk

Abstract

Metal-organic frameworks (MOFs) represent a vast chemical space of crystalline materials with diverse applications, yet existing ML-based MOF property prediction methods either use computationally intensive 3D models or rely solely on sequential processing of SMILES notation, failing to fully exploit the rich chemical and structural information inherent in SMILES text representations. In this paper, we present M-MOFormer, a novel multi-modal transformer framework that integrates both SMILES representations and automatically generated 2D structural diagrams through our developed openchemlib-MMOF package. By incorporating cross-modal attention mechanisms, M-MOFormer effectively combines structural information from textual and visual modalities, achieving superior prediction accuracy compared to existing structure-agnostic approaches. Our interpretability analysis reveals M-MOFormer's ability to identify chemically relevant structural features, particularly metal-ligand coordination sites and carbonyl groups. Additionally, we open-source a comprehensive multi-modal MOF prediction dataset, establishing new benchmarks for structure-agnostic MOF property prediction while maintaining computational efficiency.

Introduction

Metal-organic frameworks (MOFs) represent a revolutionary class of crystalline materials that have transformed materials science research over the past decades (James 2003; Zhou, Long, and Yaghi 2012). Their distinctive combination of metal nodes and organic linkers creates highly porous three-dimensional structures with unprecedented versatility in chemical composition and topology (Sharp et al. 2021), enabling widespread applications across gas absorption, water harvesting, and energy storage (Ahmed et al. 2019; Almassad et al. 2022; Cao, Liu, and Barati Farimani 2019). While the vast chemical space of possible MOF structures—arising from numerous combinations of metal nodes, organic linkers, and topological arrangements—presents both opportunities and significant challenges for materials discovery (Moosavi et al. 2020; Falcaro et al. 2011), the exponential growth in potential MOF structures has created an urgent need for efficient property prediction methods. Given that traditional experimental characterization, being both time-intensive and resource-demanding, cannot keep pace with the rapid expansion of possible MOF candidates (Wilmer et al. 2012), this limitation has spurred the development of computational approaches, with machine learning (ML) emerging as a promising direction for accelerating MOF discovery and optimization (Fung et al. 2021).

Current ML-based MOF property prediction methods primarily follow two distinct approaches. The first approach, exemplified by the Crystal Graph Convolutional Neural Network (CGCNN) (Xie and Grossman 2018), utilizes detailed three-dimensional (3D) structural models for MOF prediction. While these methods achieve superior prediction accuracy by leveraging comprehensive structural information, they face significant practical limitations. The requirement for optimized 3D atomic coordinates introduces substantial computational overhead, particularly for large MOF structures containing thousands of atoms. Additionally, the memory-intensive nature of crystal graph operations can make this approach prohibitive for large-scale screening applications (Wang et al. 2020).

The second approach prioritizes computational efficiency by utilizing simplified molecular representations, primarily SMILES (Simplified Molecular Input Line Entry System) (Weininger 1988), as demonstrated by the MOFormer model (Cao et al. 2023). The MOFid representation, combining SMILES notation with topology and catenation information, provides a concise yet informative text-based description of MOFs (Bucior et al. 2019). While these approaches enable rapid processing and scalability, they primarily extract knowledge from SMILES strings in a sequential manner without fully exploiting their inherent chemical and structural information, thus typically achieving lower prediction accuracy compared to 3D structure-based methods.

To bridge this gap, we propose M-MOFormer, a multimodal transformer framework, that leverages both SMILES representations and automatically generated 2D structural diagrams for accurate MOF property prediction. First, we introduce openchemlib-MMOF, an automated visualization python package that generates chemically accurate 2D structural representations from SMILES notations, to create a comprehensive multimodal MOF dataset. Next, a novel transformer architecture,M-MOFormer, is proposed in this work to effectively integrates complementary struc-

[*]Corresponding author

tural information from both textual and visual modalities through cross-modal attention mechanisms. Extensive experiments demonstrate that M-MOFormer significantly outperforms existing structure-agnostic approaches across multiple datasets while approaching the accuracy of 3D structure-based methods. Importantly, our interpretability analysis reveals that M-MOFormer can successfully identify chemically relevant structural features, particularly metal-ligand coordination sites and carbonyl groups, aligning with theoretical understanding of MOF properties.

The main contributions of this work are as follows:

1. We introduce M-MOFormer, an novel multimodal transformer framework, for effective structure-agnostic MOF prediction.
2. We open-source a comprehensive multimodal MOF prediction dataset that includes both SMILES expressions and corresponding 2D chemical structures.
3. The extensive experimentation demonstrate the superior performance of M-MOFormer across multiple tasks, establishing new benchmarks for structure-agnostic MOF property prediction.

Related Work

The evolution of MOF property prediction methods reflects a progressive shift from traditional machine learning approaches to sophisticated deep learning architectures. Here we present a structured overview of this development, highlighting key methodological advances and implications.

Traditional Machine Learning Approaches

Initial efforts in MOF property prediction relied on conventional machine learning algorithms such as support vector machines (SVM), random forests, and gradient boosting machines (Meredig et al. 2014). These methods typically employed hand-crafted features derived from MOF characteristics, including pore size distribution, surface area, and void fraction (Moghadam et al. 2019). While these approaches established foundational methodologies for property prediction, their effectiveness was inherently limited by the quality and completeness of manually engineered features, often failing to capture complex structural relationships.

Deep Learning Approaches

Deep learning methods have revolutionized MOF property prediction by enabling automatic feature extraction, advancing beyond the limitations of traditional machine learning approaches. This evolution encompasses several key paradigms, each with distinct advantages and challenges.

The Crystal Graph Convolutional Neural Network (CGCNN) (Xie and Grossman 2018) pioneered the use of graph-based architectures for MOF property prediction. By modeling atomic structures as graphs, these approaches effectively capture complex three-dimensional interactions (Fung et al. 2021). However, their reliance on detailed 3D structural information introduces significant computational overhead, limiting their applicability in high-throughput screening scenarios. To address computational efficiency, sequence-based approaches emerged utilizing SMILES representations (Weininger 1988) and MOFid (Bucior et al. 2019). These methods process MOFs as standardized text strings, enabling the application of natural language processing techniques. The introduction of transformer-based models like MOFormer (Cao et al. 2023) leveraged self-attention mechanisms to capture long-range dependencies while maintaining computational efficiency. However, these approaches potentially sacrifice structural information crucial for accurate property prediction.

Recent work has begun exploring the integration of multiple molecular representations to balance accuracy and efficiency. Our work advances this direction by combining SMILES representations with 2D structural information through a novel transformer architecture, achieving both the accuracy necessary for reliable property prediction and the efficiency required for large-scale screening.

Method

Overall architecture

The overall framework of M-MOFormer, schematically illustrated in Fig. 1, comprises two key modules: the molecular structure visualization module and the M-MOFormer model. The molecular structure visualization module transforms the SMILES representation of MOFs into two-dimensional structural diagrams by visualizing every secondary building units (SBUs). Subsequently, the M-MOFormer model processes both the SMILES representation and its corresponding 2D structural diagram through text and image tokenizers, respectively. These tokens are then passed through SMILES embedding and image embedding layers to generate latent representations. Finally, a Transformer encoder explores the relationship between structural expressions and chemical properties to output chemical property predictions. The detailed structure of the employed Transformer encoder is illustrated in Fig. 1 c.

The key innovation of M-MOFormer lies in its ability to leverage both textual and visual representations of MOF structures, enabling more comprehensive feature extraction and improved prediction accuracy. This dual-modality approach allows the model to capture both the precise molecular composition from SMILES representations and the spatial structural relationships from 2D visualizations.

In the following sections, we will introduce each component of M-MOFormer in detail: the molecular structure visualization module and the M-MOFormer model.

Molecular structure visualization module

We developed the openchemlib-MMOF package as an extension of the JavaScript-based openchemlib framework, enabling robust visualization of complex SMILES strings containing multiple SBUs in a Python environment. [1]. The visualized complex SMILES are formatted as a 256×256 image.

$$\textbf{Image} \in \mathbb{R}^{256\times256} = \text{openchemlib-MMOF}(\text{SMILES}) \tag{1}$$

[1] Code is available at https://github.com/IkeYang/M-MOFormer

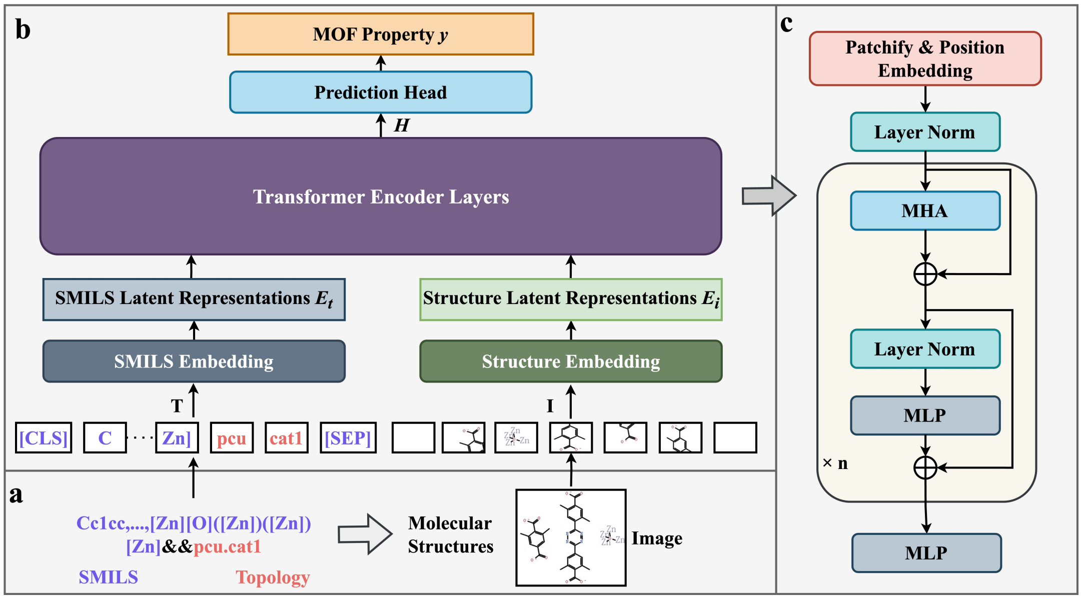


Figure 1: Overview of M-MOFormer. (a) Molecular structure visualization module to create 2D structural diagrams, (b) M-MOFormer model for multi-modal feature extraction and MOF property prediction, and (c) Transformer encoder architecture.

where SMILES $\in \mathbb{R}^{L_t}$ and L_t is the sequence lengths.

The proposed M-MOFormer model

The M-MOFormer model processes multi-modal inputs of SMILES and **Image** through a carefully designed transformer architecture that can be mathematically formulated as follows:

$$\begin{aligned}
\mathbf{T} \in \mathbb{R}^{L_t \times d_t} &= \text{TextTokenizer}(\text{SMILES}) \\
\mathbf{I} \in \mathbb{R}^{L_i \times d_i} &= \text{ImageTokenizer}(\text{Image}) \\
\mathbf{E}_t \in \mathbb{R}^{L_t \times d} &= \text{SMILESEmbedding}(\mathbf{T}) \\
\mathbf{E}_i \in \mathbb{R}^{L_i \times d} &= \text{ImageEmbedding}(\mathbf{I}) \\
\mathbf{H} \in \mathbb{R}^{(L_t+L_i) \times d} &= \text{TransformerEncoder}([\mathbf{E}_t; \mathbf{E}_i]) \\
\mathbf{y} \in \mathbb{R} &= \text{MLP}(\mathbf{H})
\end{aligned} \tag{2}$$

where L_i is the sequence lengths of image tokens; d_t and d_i are the initial token dimensions; d is the model hidden dimension; and p is the number of predicted properties.

Computational experiments

Experimental Setup

We evaluate our proposed method on two widely-used MOF benchmark datasets: **1) Quantum MOF (QMOF)** (Rosen et al. 2021): A dataset comprising 20,375 MOFs with DFT-calculated band gaps (in eV) as property labels. **2) hypothetical MOFs (hMOF)** (Wilmer et al. 2012): A collection of 137,652 MOFs with gas adsorption properties. The dataset provides CO_2 and CH_4 uptake capacities (in mol·kg^{-1}) measured at three pressure points: 0.05, 0.5, and 2.5 bar.Following the standard protocol (Cao et al. 2023), we randomly partition each dataset into training (70%), validation (15%), and test (15%) sets.

For comprehensive evaluation, we compare our method against two categories of baseline approaches: **1) Structure-dependent methods:** These methods leverage detailed 3D atomic structures of MOFs for property prediction, typically achieving higher accuracy due to the rich structural information. Representative approaches include CGCNN (Xie and Grossman 2018) and Smooth Overlap of Atomic Position (SOAP) (Himanen et al. 2020). **2) Structure-agnostic methods:** These methods operate without requiring explicit structural information, relying primarily on molecular representations such as SMILES. State-of-the-art approaches in this category include MOFormer (Cao et al. 2023) and Stoichiometric-120 (Meredig et al. 2014).

To ensure fair comparison, we adopt the same transformer architecture as (Cao et al. 2023), with the following specifications: vocabulary size of 4021 tokens, d_t and d_i equals 512 and 50 respectively, 8 attention heads, hidden dimension of 512, 6 transformer layers, and dropout rate of 0.1.

Main Results

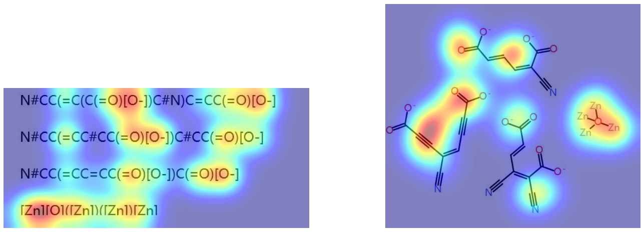
Table 1 presents the performance comparison between our proposed M-MOFormer and baseline methods across both QMOF and hMOF datasets. Among structure-dependent methods, SOAP achieves superior performance on hMOF dataset with the lowest MAE across all gas adsorption predictions, while CGCNN shows the best performance for QMOF band gap prediction with an MAE of 0.275 eV. Within structure-agnostic methods, our M-MOFormer demonstrates consistent improvements over existing approaches. Compared to MOFormer, M-MOFormer reduces MAE by 7.2% (from 0.387 to 0.359) on band gap prediction, and shows 8.3-11% error reductions for gas uptake predic-

Table 1: Benchmark performance comparison on QMOF and hMOF datasets

Model	QMOF	hMOF					
	Band gap	CO_2 0.05bar	CO_2 0.5bar	CO_2 2.5bar	CH_4 0.05bar	CH_4 0.5bar	CH_4 2.5bar
Structure-dependent methods:							
CGCNN	**0.275**	0.126	0.391	0.818	0.028	0.121	0.333
SOAP	0.424	**0.115**	**0.339**	**0.666**	**0.022**	**0.106**	**0.239**
Structure-agnostic methods:							
M-MOFormer	**0.359**	**0.169**	**0.504**	**0.889**	**0.031**	**0.158**	**0.343**
MOFormer	0.387	0.178	0.558	1.000	0.034	0.174	0.385
Stoichiometric-120	0.466	0.282	0.983	1.895	0.050	0.269	0.631

Table 2: Ablation analysis of considered molecular representations of M-MOFormer for CO_2 uptake prediction at 0.05 bar in hMOF.

Components		**Performance**
SMILES	Structural diagrams	MAE
✓	✗	0.172
✗	✓	0.280
✓	✓	0.169


N#CC(=C(C(=O)[O-])C#N)C=CC(=O)[O-]
N#CC(=CC#CC(=O)[O-])C#CC(=O)[O-]
N#CC(=CC=CC(=O)[O-])C(=O)[O-]
[Zn][O]([Zn])([Zn])[Zn]

(a) SMILES attention

(b) 2D structure attention

Figure 2: GradCAM visualization of M-MOFormer's attention patterns. (a) SMILES representation and (b) 2D molecular structure.

tions. The improvements are more substantial when compared to Stoichiometric-120, with 22.9% reduction in band gap prediction and up to 53.1% reduction in gas adsorption predictions. While structure-dependent methods maintain their advantage in absolute performance, M-MOFormer significantly narrows this gap while maintaining the efficiency of structure-agnostic approaches.

Ablation Studies

To investigate the effectiveness of different molecular representations in M-MOFormer, we conduct ablation experiments on the hMOF dataset for CO_2 uptake prediction at 0.05 bar. As shown in Table 2, using SMILES representation alone achieves an MAE of 0.172 $\text{mol}\cdot\text{kg}^{-1}$, significantly outperforming the structural diagram-only variant (MAE = 0.280 $\text{mol}\cdot\text{kg}^{-1}$). This indicates that SMILES encoding captures more essential molecular information for property prediction. When combining both representations, M-MOFormer achieves the best performance with an MAE of 0.169 $\text{mol}\cdot\text{kg}^{-1}$, demonstrating a modest but consistent improvement over the SMILES-only variant. These results validate our design choice of incorporating dual molecular representations, where SMILES serves as the primary feature extractor while structural diagrams provide complementary information for enhanced prediction accuracy.

Interpretability analysis

To provide insights into how M-MOFormer makes predictions, we visualize the attention patterns using GradCAM (Selvaraju et al. 2017) for both SMILES and 2D structural representations, as shown in Fig. 2.

The attention visualizations reveal that M-MOFormer primarily focuses on metal-ligand coordination sites and C=O double bonds. This observation aligns with theoretical MOF studies - carbonyl groups typically exhibit strong adsorption energies due to their polar nature and electron-rich characteristics. The consistent attention patterns across both modalities demonstrate that M-MOFormer successfully learns to identify chemically relevant structural features that determine MOF properties.

Conclusion

This work introduced M-MOFormer, a novel multi-modal transformer framework for advances MOF property prediction. By integrating automatically generated 2D structural diagrams through our openchemlib-MMOF package and employing cross-modal attention mechanisms with transformer architecture, our approach could better mine the rich chemical and structural information inherent in SMILES representations.

Experimental results demonstrated that M-MOFormer significantly outperformed existing structure-agnostic methods across multiple prediction tasks, achieving SOTA prediction accuracy. While structure-dependent methods maintained a slight advantage in absolute performance, M-MOFormer approached their accuracy without requiring computationally expensive 3D structural information. Interpretability analysis revealed that our model successfully identified chemically relevant structural features, particularly focusing on metal-ligand coordination sites and carbonyl groups, which aligned with theoretical understanding of MOF properties.

Acknowledgments

This research is supported by Shaanxi Province Key Projects (103-428824034).

References

Ahmed, A.; Seth, S.; Purewal, J.; Wong-Foy, A. G.; Veenstra, M.; Matzger, A. J.; and Siegel, D. J. 2019. Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nature Communications*, 10: 1–9.

Almassad, H. A.; Abaza, R. I.; Siwwan, L.; Al-Maythalony, B.; and Cordova, K. E. 2022. Environmentally adaptive MOF-based device enables continuous self-optimizing atmospheric water harvesting. *Nature Communications*, 13(1): 1–10.

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

Cao, Z.; Liu, V.; and Barati Farimani, A. 2019. Water desalination with two-dimensional metal-organic framework membranes. *Nano Letters*, 19(12): 8638–8643.

Cao, Z.; Magar, R.; Wang, Y.; et al. 2023. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *Journal of the American Chemical Society*, 145(5): 2958–2967.

Falcaro, P.; Hill, A. J.; Nairn, K. M.; Jasieniak, J.; Mardel, J. I.; Bastow, T. J.; Mayo, S. C.; Gimona, M.; Gomez, D.; Whitfield, H. J.; et al. 2011. A new method to position and functionalize metal-organic framework crystals. *Nature Communications*, 2(1): 1–8.

Fung, V.; Zhang, J.; Juarez, E.; and Sumpter, B. G. 2021. Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1): 1–8.

Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; and Foster, A. S. 2020. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247: 106949.

James, S. L. 2003. Metal-organic frameworks. *Chemical Society Reviews*, 32: 276–288.

Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; and Wolverton, C. 2014. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, 89(9): 094104.

Moghadam, P. Z.; Rogge, S. M.; Li, A.; Chow, C.-M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V.; et al. 2019. Structure-mechanical stability relations of metal-organic frameworks via machine learning. *Matter*, 1(1): 219–234.

Moosavi, S. M.; Nandy, A.; Jablonka, K. M.; Ongari, D.; Janet, J. P.; Boyd, P. G.; Lee, Y.; Smit, B.; and Kulik, H. J. 2020. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1): 1–10.

Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; and Snurr, R. Q. 2021. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5): 1578–1597.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Sharp, C. H.; Bukowski, B. C.; Li, H.; Johnson, E. M.; Ilic, S.; Morris, A. J.; Gersappe, D.; Snurr, R. Q.; and Morris, J. R. 2021. Nanoconfinement and mass transport in metal-organic frameworks. *Chemical Society Reviews*, 50: 11530–11581.

Wang, R.; Zhong, Y.; Bi, L.; Yang, M.; and Xu, D. 2020. Accelerating Discovery of Metal-Organic Frameworks for Methane Adsorption with Hierarchical Screening and Deep Learning. *ACS Applied Materials & Interfaces*, 12(47): 52797–52807.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31–36.

Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; and Snurr, R. Q. 2012. Large-scale screening of hypothetical metal-organic frameworks. *Nature Chemistry*, 4(2): 83–89.

Xie, T.; and Grossman, J. C. 2018. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14): 145301.

Zhou, H.-C.; Long, J. R.; and Yaghi, O. M. 2012. Introduction to metal-organic frameworks. *Chemical Reviews*, 112: 673–674.