MULTIMODAL DOMAIN-ADAPTIVE LANGUAGE MODELS FOR MATERIALS DISCOVERY IN METAL–ORGANIC FRAMEWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate and data-efficient prediction of metal–organic framework (MOF) properties requires representations that capture both chemical and structural information. We propose a two-stage framework that continually pretrains the Granite language model on PubChem records and MOF literature, and then fuses its frozen MOFid representations with 18 interpretable CIF-derived descriptors through cross-attention. Continued pretraining alone improves QMOF band-gap and HMOF gas-adsorption prediction, and structural grounding adds further gains. Controlled ablations show that combining pretrained MOFid language representations with structural descriptors outperforms using either the pretrained MOFid representations or the structural descriptors alone, while cross-attention generally performs better than concatenation under matched conditions. Representation strengths are property-specific: structurally pretrained models perform best on QMOF and CH_4, whereas our multimodal framework achieves the best results on all three CO_2 targets. Bond Connectivity descriptors are most effective for QMOF, and Pore Structure descriptors are effective to HMOF. The two modalities remain complementary in low-data settings, demonstrating that compact structural grounding can effectively enhance MOFid-based language representations.

1 INTRODUCTION

Metal-organic frameworks (MOFs) are in general porous crystalline materials constructed from combinations of metal nodes, organic linkers, and network topologies (James, 2003; Zhou et al., 2012). Their vast combinatorial design space makes them promising candidates for applications in carbon capture, gas storage, and catalysis (Boyd et al., 2019; Ahmed et al., 2019; Bavykina et al., 2020). Since this design space is far too large to explore exhaustively through simulation or synthesis, predicting material properties directly from structure has become a practical strategy for high-throughput MOF screening (Wilmer et al., 2012; Moosavi et al., 2020b). A central bottleneck, however, is representation: determining what structural information a model must receive to predict a target property accurately, as well as the computational or experimental cost of generating that input.

Therefore, data-driven approaches, including machine learning (ML), have emerged as powerful tools for accelerating MOF discovery and understanding structure–property relationships (Wilmer et al., 2012; Boyd et al., 2019). A central challenge in applying ML to MOFs is developing representations that are both informative and computationally efficient. Structure-based models, such as Crystal Graph Convolutional Neural Networks (CGCNNs), encode periodic atomic environments, whereas the Porous Materials Transformer (PMTransformer) combines atom-based graph embeddings with three-dimensional interaction-energy grids to capture local chemical environments and global pore characteristics (Xie & Grossman, 2018a; Park et al., 2023a). Although these models achieve strong predictive performance, they require detailed atomistic or spatial representations of the complete crystal structure. Such representations can increase computational cost for MOFs with large unit cells and are generally less directly interpretable than compact, physically meaningful descriptors. These limitations motivate the development of compact representations that preserve relevant chemical and structural information without requiring full atomistic graphs or spatial-grid encodings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To reduce the computational demands associated with full atomistic representations, compact text-based representations have emerged as an efficient alternative. In particular, MOFid represents MOFs as strings that combine the Simplified Molecular Input Line Entry System (SMILES) representations of their molecular building blocks with high-level structural information, such as topology and catenation (Bucior et al., 2019; Weininger, 1988; O'Keeffe et al., 2008). This string-based representation allows Transformer models to process MOFs directly, as demonstrated by MOFormer (Cao et al., 2023a). MOFormer uses a Transformer encoder to learn representations from MOFid strings and aligns them with crystal-graph representations through self-supervised pretraining. However, it remains unclear when compact MOFid representations are sufficient for property prediction and when explicitly integrating them with structural descriptors provides complementary information.

To provide explicit structural grounding for MOFid-based language models, descriptors derived from Crystallographic Information Files (CIFs) offer a lightweight alternative to full atomistic neural encoders by summarizing physically meaningful aspects of MOF structures. Although these descriptors are widely used as standalone inputs, their complementarity with representations learned from MOFid strings remains underexplored. Rather than treating MOFid strings, handcrafted descriptors, and structure-aware neural encoders as separate alternatives, we investigate whether lightweight structural grounding can improve compact MOFid-based language representations and how much explicit structural information is required for accurate property prediction. We evaluate this complementarity on two sets of tasks: QMOF band-gap prediction and HMOF CH_4 and CO_2 gas-adsorption prediction at three pressure conditions (Wilmer et al., 2012; Rosen et al., 2021a; 2022).

Research Questions. We investigate how compact physical information can ground MOFid representations for property prediction through five research questions:

- **RQ1: Impact of domain-adaptive continued pretraining.** To what extent does continued pretraining of a general-domain language model on PubChem records and MOF-related literature improve QMOF and HMOF property prediction, particularly in low-data regimes?
- **RQ2: Structural complementarity.** Can compact CIF-derived descriptors complement frozen MOFid representations and achieve performance competitive with that of dense atomistic structure encoders?
- **RQ3: Cross-modal interaction.** Does descriptor-guided cross-attention between MOFid tokens and physical descriptors outperform standard concatenation of their pooled representations?
- **RQ4: Property-specific structural complementarity.** Which CIF-derived descriptor families most effectively complement MOFid representations for QMOF band-gap and HMOF gas-adsorption prediction?
- **RQ5: Data-efficient multimodal transfer.** Can a language model pretrained on MOF literature and subsequently fused with compact CIF-derived descriptors outperform dense atomistic structure encoders pretrained on substantially larger MOF and porous-material structure datasets when downstream labels are scarce?

To answer these questions, we conduct controlled experiments on the QMOF band-gap and HMOF gas-adsorption tasks. We compare the original Granite model with its domain-adapted variant to examine how continued pretraining affects MOFid-based property prediction. We then investigate structural grounding by integrating frozen MOFid token representations obtained from the continually pretrained Granite model with 18 physically interpretable CIF-derived descriptors using descriptor-guided cross-attention. We compare our proposed models with MOFid-based, descriptor-based, and dense atomistic baselines. We also evaluate standard concatenation, descriptor-family ablations, compact descriptor subsets, modality-corruption experiments, and low-data-regime experiments. These controlled experiments determine when lightweight structural grounding improves MOFid-based prediction, how the fusion mechanism affects multimodal learning, which physical signals provide the greatest benefit across properties, and how our proposed models perform against the baselines under limited supervision.

Contributions. Our main contributions are:

- We introduce a two-stage framework that first continually pretrains a general-domain Granite language model on PubChem records and open-access MOF-related abstracts and then

integrates the resulting frozen MOFid representations with 18 compact CIF-derived descriptors through descriptor-guided cross-attention.

- We conduct controlled modality and fusion experiments showing that MOFid representations and physical descriptors provide complementary predictive signals. Neither the MOFid-only model nor the descriptor-only model matches the multimodal model. The fusion comparison further shows that cross-attention generally yields better predictive performance than standard concatenation under matched inputs and training conditions.

- We identify property-specific structural requirements through descriptor-family and compact-subset ablations. Bond Connectivity provides the strongest complement to MOFid representations for QMOF band-gap prediction, whereas HMOF adsorption benefits most from Pore Structure descriptors. Notably, accessible volume and accessible surface area alone achieved strong predictive performance across the gas adsorption tasks.

- We demonstrate effective transfer under limited supervision against a dense atomistic model pretrained on 1.9 million porous-material structures. With only 500 labeled examples, our model leads on QMOF, most CH_4 targets, all CO_2 targets, and overall HMOF performance, while retaining its advantages on QMOF and all CO_2 targets and remaining competitive on CH_4 targets at 2,000 examples.

2 Related Work

Domain-adaptive continued pretraining has improved language models in scientific, biomedical, materials, and chemical domains (Lee et al., 2020; Beltagy et al., 2019; Gupta et al., 2022; Chithrananda et al., 2020; Ross et al., 2022). However, its role in MOF property prediction remains underexplored. Existing MOF models primarily learn from MOFid representations, align MOFid representations with structural encoders, or rely heavily on pretraining with structural datasets. This leaves the potential impact of continued pretraining on molecular records and MOF-focused literature for MOF property prediction underexplored.

Descriptor-based approaches represent MOFs using engineered geometric, chemical, and topological features and apply nonlinear models such as Random Forest and XGBoost to predict adsorption and electronic properties (Fernandez et al., 2014; Rosen et al., 2021b; Moosavi et al., 2020a). SOAP represents local atomic environments by expanding Gaussian-smoothed atomic densities using radial basis functions and spherical harmonics, producing structural fingerprints invariant to rotations, translations, and permutations of equivalent atoms (Bartók et al., 2013). These approaches demonstrate the predictive value and interpretability of physically motivated representations, but encode each MOF as a fixed feature vector. Consequently, they do not model interactions between individual structural descriptor families and token-level chemical representations.

Recent MOF learning methods span symbolic and atomistic representations. CGCNN represents periodic crystals as atomic graphs and learns local structural features through message passing between atoms and their neighbors (Xie & Grossman, 2018b). MOFormer instead processes MOFid strings encoding framework building blocks and topology. It acquires structural information through self-supervised alignment with CGCNN representations derived from the corresponding crystal structures, but requires only MOFid during downstream prediction (Cao et al., 2023b).

MOFTransformer combines atom-based graph embeddings with three-dimensional energy-grid embeddings to represent local chemistry and global pore environments, and is pretrained on about 1.1 million hypothetical MOFs (Kang et al., 2023). PMTransformer extends this framework through pretraining on about 1.9 million hypothetical porous materials, including MOFs, covalent-organic frameworks, porous polymer networks, and zeolites (Park et al., 2023b). Together, these approaches demonstrate the value of atomic environments, MOFid semantics, crystal graphs, and large-scale structural pretraining. However, they do not determine which compact physical descriptor families best complement MOFid, whether domain-adapted MOFid representations remain useful after structural grounding, or how physical descriptors should interact with individual MOFid tokens. Answering these questions could advance physically grounded MOF textual representations.

3 METHODOLOGY

We propose a two-stage framework for MOF property prediction that integrates domain-adapted language modeling with crystal structural information. Our approach proceeds in two phases: (i) continued domain-adaptive pretraining of the Granite 3.3 2B model (IBM Granite Team, 2025) on chemistry-focused corpora, and (ii) construction of a multimodal architecture that fuses MOFid textual representations with CIF-derived structural descriptors through descriptor-guided cross-attention.

3.1 TASK DEFINITION AND DATASETS

We evaluate the framework on quantum-property prediction using QMOF and gas-adsorption prediction using HMOF. We adopt the same dataset splits used by MOFormer (Cao et al., 2023a).

HMOF. The HMOF dataset contains gas-adsorption properties for CO_2 and CH_4 in $\mathrm{mol\,kg^{-1}}$ at pressures of 0.05, 0.5, and 2.5 bar. Of the 137,652 MOFs in the original dataset, 102,858 have MOFid representations. The training, validation, and test splits contain 72,000, 15,428, and 15,428 samples, respectively. Each sample contains a MOFid and six adsorption targets.

QMOF. The QMOF dataset contains quantum-chemical properties for 20,375 MOF structures, of which 7,466 have MOFid representations. We predict the PBE band gap in electron volts (Canepa et al., 2013). The training, validation, and test splits contain 5,226, 1,119, and 1,119 samples, respectively.

3.2 DOMAIN-ADAPTIVE CONTINUED PRETRAINING

To improve the representation of chemical notations such as SMILES and MOFid, we perform continued pretraining of Granite on chemistry-focused corpora. We consider two pre-training setups: (i) PubChem alone, using 250K compound records totaling 254 MB, and (ii) PubChem followed by open-access MOF-related articles and abstracts from PubMed Central, resulting in a combined corpus size of 1.13 GB (Kim et al., 2024; PMC).

Each PubChem record combines the compound name with available chemical information, including its SMILES representation, molecular formula, and molecular weight. We also collected a MOF-related corpus from the PMC Open Access Subset using MOF-specific search terms, yielding 28,525 articles and abstracts. The abstracts are segmented into sentences using NLTK (Bird & Loper, 2004). Our pre-training phase uses a causal language-modeling objective and does not include downstream property labels or CIF-derived descriptors. Further details on corpus construction and preprocessing are provided in Appendix A, while the leakage analysis is reported in Appendix B.

3.3 CIF-DERIVED STRUCTURAL DESCRIPTORS

We use the CIF files corresponding to the 102,858 HMOF and 7,466 QMOF samples with MOFid representations. From each CIF, we extract 18 compact structural descriptors organized into four groups: Metal–Ligand Geometry, Metal Coordination, Bond Connectivity, and Pore Structure. Metal neighbors, coordination environments, and structure graphs are obtained using CrystalNN as implemented in PYMATGEN (Pan et al., 2021; Ong et al., 2013).

Metal–Ligand Geometry. Minimum, mean, and maximum metal–ligand distances describe the central tendency and range of local metal–ligand geometries within each framework.

Metal Coordination. The minimum, maximum, mean, and standard deviation of coordination numbers across metal sites capture the typical coordination environment, its range, and the degree of coordination heterogeneity within the framework.

Bond Connectivity. Counts of carbon–oxygen, metal–oxygen, carbon–carbon, metal–nitrogen, and carbon–nitrogen connections summarize connection-type frequencies within the framework.

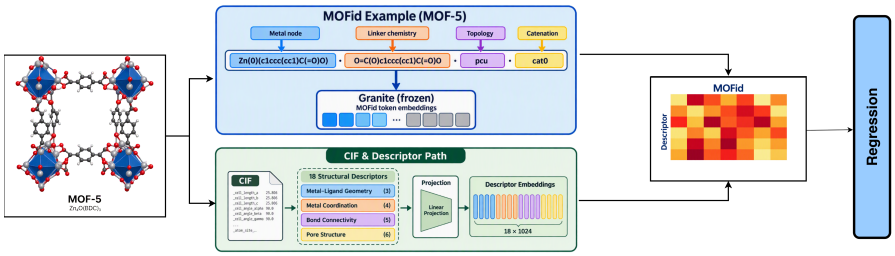


Figure 1: Overview of the proposed multimodal framework. CIF-derived structural descriptors guide cross-attention over frozen MOFid token representations, and the fused representation is passed to a regression head.

Pore Structure. We use six lightweight proxy descriptors to represent pore-related structural information: accessible volume, accessible surface area, largest cavity diameter, largest free sphere, pore-limiting diameter, and void fraction. These descriptors serve as structural conditioning features in the multimodal model and should not be interpreted as rigorous probe-accessible or experimentally measured pore properties.

Complete feature definitions and extraction procedures are reported in Appendix C.

3.4 Descriptor-Guided Multimodal Fusion

We combine frozen Granite token embeddings with CIF-derived descriptors through descriptor-guided cross-attention. As shown in Figure 1, the normalized 18-dimensional descriptor vector is projected into a 1024-dimensional fusion space and used as the query, while the MOFid token embeddings provide the keys and values. This allows the structural descriptors to attend to the MOFid tokens that are most relevant to the physical context they encode.

Let $H \in \mathbb{R}^{L \times d_h}$ denote the frozen MOFid token representations and $x \in \mathbb{R}^{18}$ denote the normalized descriptor vector. Trainable projection matrices produce the query $q = W_q x$, keys $K = HW_k$, and values $V = HW_v$. Descriptor-guided cross-attention is then computed as

$$z = \mathrm{softmax}\left(\frac{qK^\top}{\sqrt{d}}\right) V,$$

where $d = 1024$ is the attention dimension.

The fusion module consists of six transformer layers with 16 attention heads and a feed-forward dimension of 4096. The fused representation z is passed to a regression head for property prediction. During fine-tuning, the Granite backbone remains frozen, while the descriptor projection, fusion module, and regression head are trained. This configuration introduces approximately 78M trainable parameters in addition to the frozen 2B-parameter Granite backbone.

3.5 Experimental Protocol

The evaluation covers three settings: generative prediction from MOFid, LLM-only regression, and multimodal regression. All LLM-based models use the Granite-3.3-2B model. For multimodal prediction, the Granite backbone is frozen, while the descriptor projection, fusion module, and regression head are trained. To assess the effect of the fusion mechanism, we compare descriptor-guided cross-attention against two concatenation variants using the same frozen backbone and complete descriptor set. In cross-attention, the descriptor vector queries the full MOFid token sequence, allowing the model to selectively attend to relevant tokens. Concat-FFN replaces the attention sub-layer with six FFN layers of identical depth and width, making the attention mechanism the primary architectural difference. Concat-Linear uses a single projection layer after merging the two modalities, serving as a lightweight baseline. Full architectural details are reported in Appendix D.

We compare our proposed model against Random Forest, XGBoost, SOAP, CGCNN, MOFormer, MOFTransformer, and PMTransformer, covering descriptor-based, atomic-environment, crystal-graph, MOFid-based, and structure-pretrained representations. For SOAP, CGCNN, and MOFormer, we report the results from the original MOFormer study (Cao et al., 2023b), which evaluates each model over three runs. MOFTransformer, PMTransformer, Random Forest, and XGBoost are reproduced using the official PMTransformer code and hyperparameters (Kang et al., 2023; Park et al., 2023b). For MOFTransformer and PMTransformer, we use a unified regression head across the six HMOF targets, as our preliminary evaluation showed performance comparable to separate target-specific models. Random Forest and XGBoost use the same 18 CIF-derived descriptors as our framework. All experiments in our study use the same QMOF and HMOF dataset splits adopted by MOFormer (Cao et al., 2023b).

Moreover, to evaluate the generalization and computational efficiency of our approach across backbone families and model scales, we construct two additional multimodal variants using Granite 4.0 350M (Granite Team, IBM, 2025) and Phi-4 Mini (Microsoft et al., 2025). Furthermore, to evaluate performance under limited supervision, we randomly sample 500 and 2K examples from each original training split using a fixed seed, while retaining the original validation and test sets. Within each data regime, all models use identical training subsets and are trained on these subsets for up to 100 epochs to provide sufficient optimization steps given the small number of batches per epoch.

Across both full-data and limited-supervision settings, we evaluate QMOF and each HMOF target using mean absolute error averaged over three runs with seeds 42, 123, and 1234. HMOF Mean denotes the arithmetic mean of the six target-level MAEs. For external baselines, we preserve the numerical precision of the originally reported results. Standard deviations, individual-run results, hyperparameters, and computational details are provided in Appendices G and E.

4 Results and Discussion

We evaluate the impact of domain-adaptive continued pretraining, descriptor grounding, cross-modal fusion, and limited supervision. Table 1 presents the full-data comparison, Table 2 isolates the contributions of each modality and descriptor family, and Table 12 reports results when models are fine-tuned on 500 and 2K examples only.

4.1 Domain-Adaptive Continued Pretraining

RQ1: Does domain-adaptive pretraining improve MOF property prediction?

Table 1 shows that domain-adaptive continual pretraining improves QMOF and HMOF prediction in both the generative and multimodal settings. For generative models, QMOF MAE decreases from 0.3513 for the general-domain Granite model to 0.3334 using the PubChem corpus and 0.3130 using the combined corpus of PubChem records and MOF-related articles and abstracts. HMOF Mean improves more modestly from 0.2549 to 0.2522 and 0.2508, respectively, with the largest target-level improvement occurring for CO_2 adsorption at 2.5 bar. This smaller aggregate improvement may reflect the larger HMOF training set, which contains 72,000 examples compared with 5,226 for QMOF, reducing the relative contribution of continual pretraining when more supervised data are available.

In the multimodal setting, incorporating all four descriptor groups with the adapted backbone reduces QMOF MAE from 0.3031 to 0.2878 and HMOF Mean from 0.2273 to 0.2213. The largest HMOF improvement is again observed for CO_2 adsorption at 2.5 bar, where MAE decreases from 0.5589 to 0.5450. Overall, continual pretraining provides its clearest benefits for QMOF and high-pressure CO_2 adsorption. To assess whether continual pretraining becomes more important when fewer supervised examples are available, we conduct limited-supervision experiments using 500 and 2K training examples in Section 4.5.

4.2 Comparison with Structure-Based Models

RQ2: Can compact structural descriptors complement MOFid and compete with dense atomistic encoders?

Table 1: Full-data test MAE on QMOF and six HMOF tasks, averaged over three seeds. HMOF Mean is the arithmetic mean across the six adsorption tasks. Parentheses in backbone names indicate the pretraining data. Standard deviations and individual-run results are reported in Appendix G. Lower is better; bold and underlined values indicate the best and second-best means, respectively.

Backbone	Configuration	QMOF	HMOF: CH_4			HMOF: CO_2			HMOF
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5	Mean
External baselines									
Random Forest	CIF Descriptors	0.5117	0.0248	0.1386	0.3439	0.1446	0.4492	0.7667	0.3113
XGBoost	CIF Descriptors	0.5181	0.0278	0.1543	0.3875	0.1570	0.4962	0.8607	0.3473
SOAP	Atomic Environment	0.4240	0.0220	0.1060	0.2390	0.1150	0.3390	0.6660	0.2478
CGCNN	Scratch	0.2750	0.0280	0.1210	0.3330	0.1260	0.3910	0.8180	0.3028
CGCNN (400K MOF Structures)	Pretrained	0.2560	0.0250	0.0990	0.2580	0.1100	0.3300	0.6450	0.2445
MOFormer	Scratch	0.3870	0.0340	0.1740	0.3850	0.1780	0.5580	1.0000	0.3882
MOFormer (400K MOFid–Structure)	Pretrained	0.3670	0.0330	0.1610	0.3840	0.1580	0.5450	0.9820	0.3772
MOFTransformer (1.1M MOF Structures)	Pretrained	0.3684	0.0181	0.0977	0.2409	0.1132	0.3601	0.6710	0.2501
PMTransformer (1.9M Porous Structures)	Pretrained	0.2692	0.0174	0.0896	0.2248	0.1103	0.3394	0.6406	0.2370
Our LLM models									
Granite 3.3-2B	Regression	0.4054	0.0313	0.1485	0.3432	0.1718	0.4872	0.8445	0.3378
Granite 3.3-2B	Generative	0.3513	0.0193	0.1026	0.2481	0.1220	0.3724	0.6649	0.2549
Granite 3.3-2B + PubChem	Generative	0.3334	0.0191	0.1020	0.2467	0.1206	0.3676	0.6573	0.2522
Granite 3.3-2B + PubChem + Articles	Generative	0.3130	0.0192	0.1023	0.2465	0.1201	0.3661	0.6504	0.2508
Our multimodal models									
Granite 3.3-2B + PubChem + Articles	All Descriptors	0.2878	0.0190	0.0978	0.2319	0.1086	0.3252	0.5450	0.2213
Granite 3.3-2B + PubChem + Articles	Bond Connectivity	0.2796	0.0202	0.1046	0.2467	0.1145	0.3488	0.6115	0.2411
Granite 3.3-2B + PubChem + Articles	Pore Structure	0.2970	0.0192	0.0984	0.2319	0.1092	0.3262	0.5477	0.2221
Granite 3.3-2B + PubChem + Articles	Bonding + Pore	0.2862	0.0189	0.0974	0.2298	0.1083	0.3260	0.5447	0.2209
Backbone and scale comparisons without CPT									
Granite 4.0-350M	All Descriptors	0.3364	0.0202	0.1044	0.2476	0.1146	0.3452	0.5846	0.2361
Granite 3.3-2B	All Descriptors	0.3031	0.0197	0.1010	0.2379	0.1124	0.3336	0.5589	0.2273
Phi-4 Mini 4B	All Descriptors	0.3122	0.0203	0.1039	0.2443	0.1141	0.3422	0.5777	0.2338

Table 1 shows property-dependent advantages across representations. On QMOF, pretrained CGCNN and PMTransformer achieve the lowest MAEs of 0.2560 and 0.2692, respectively, while our proposed model using Bond Connectivity obtains 0.2796. Dense atomistic models therefore retain an advantage for band-gap prediction, although our proposed multimodal model remains competitive.

Comparing the scratch and pretrained CGCNN variants further demonstrates the benefit of structural pretraining, reducing QMOF MAE from 0.2750 to 0.2560. MOFTransformer, pretrained specifically on approximately 1.1 million MOF structures, obtains a QMOF MAE of 0.3684, whereas PMTransformer, pretrained on approximately 1.9 million structures spanning multiple porous-material classes, achieves 0.2692. This comparison suggests that pretraining scale and material diversity may influence transfer to QMOF, although their individual effects cannot be isolated.

For HMOF, our proposed model using Bonding + Pore achieves the numerically lowest mean of 0.2209, compared with 0.2370 for PMTransformer, 0.2445 for pretrained CGCNN, and 0.2501 for MOFTransformer. However, the target-level pattern remains property-dependent: PMTransformer performs best on all three CH_4 targets, whereas our proposed models perform best on all three CO_2 targets. Overall, structure-pretrained models are strongest for QMOF and CH_4, while our proposed descriptor-grounded MOFid models are strongest for CO_2 and achieve the numerically lowest aggregate HMOF result.

When the same CIF-derived descriptors are used without MOFid or multimodal fusion, Random Forest and XGBoost obtain QMOF MAEs of 0.5117 and 0.5181 and HMOF Means of 0.3113 and 0.3473, respectively. These results indicate that the gains achieved by our proposed multimodal model do not arise solely from the CIF-derived descriptors.

The backbone comparisons in the last three rows of Table 1 show that our multimodal approach generalizes across model scales and architectures. Granite 4.0-350M remains competitive with Granite 3.3-2B on QMOF (0.3364 vs. 0.3031) and HMOF (0.2361 vs. 0.2273), while achieving over

Table 2: Descriptor and fusion ablations using the frozen Granite backbone continued-pretrained on PubChem records and MOF-related PMC abstracts. All results are averaged over three independent runs. LLM-only uses regression heads without the other modality. AV and ASA denote accessible volume and accessible surface area, respectively.

Input Configuration	Fusion / Head	QMOF	HMOF: CH_4			HMOF: CO_2			HMOF
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5	Mean
Metal–Ligand	Cross-Attention	0.2847	0.0204	0.1055	0.2503	0.1152	0.3514	0.6169	0.2433
Metal Coordination	Cross-Attention	0.2883	0.0203	0.1054	0.2498	0.1150	0.3523	0.6166	0.2432
Bond Connectivity	Cross-Attention	0.2796	0.0202	0.1046	0.2467	0.1145	0.3488	0.6115	0.2411
Pore Structure	Cross-Attention	0.2970	0.0192	0.0984	0.2319	0.1092	0.3262	0.5477	0.2221
– AV + ASA	Cross-Attention	–	0.0191	0.0982	0.2331	0.1087	0.3267	0.5514	0.2229
Metal–Ligand + Coordination	Cross-Attention	0.2831	0.0204	0.1061	0.2499	0.1155	0.3529	0.6205	0.2442
Metal–Ligand + Coordination + Bonding	Cross-Attention	0.2803	0.0198	0.1021	0.2397	0.1130	0.3400	0.5881	0.2338
Bonding + Pore	Cross-Attention	0.2862	0.0189	0.0974	0.2298	0.1083	0.3260	0.5447	0.2209
– Bonding + AV + ASA	Cross-Attention	0.2858	0.0188	0.0973	0.2310	0.1085	0.3259	0.5464	0.2213
All Descriptors	Concat-Linear	0.5072	0.0245	0.1185	0.2741	0.1299	0.3689	0.6279	0.2573
All Descriptors	Concat-FFN	0.4760	0.0210	0.1027	0.2399	0.1122	0.3301	0.5512	0.2262
All Descriptors	Cross-Attention	0.2878	0.0190	0.0978	0.2319	0.1086	0.3252	0.5450	0.2213
– Full MOFid shuffle	Cross-Attention	1.0983	0.0619	0.3572	0.9147	0.3497	1.1814	2.3780	0.8738
– MOFid Catenation shuffle	Cross-Attention	0.2884	0.0268	0.1564	0.4147	0.1579	0.5827	1.2464	0.5627

2× faster fine-tuning and inference than both ∼78M-parameter models, the Granite 3.3-2B-based multimodal model and PMTransformer. On the other hand, despite using a different backbone architecture, Phi-4 Mini 4B achieves competitive results on both QMOF (0.3122) and HMOF (0.2325), confirming cross-architecture generalization.

4.3 Modality Complementarity and Fusion

RQ3: Which fusion strategy performs better: cross-attention or concatenation?

The LLM-only regression model, as shown in Table 1, obtains QMOF and HMOF Mean MAEs of 0.4054 and 0.3378, respectively. Integrating the complete descriptor set through cross-attention, as shown in Table 2, reduces these MAEs to 0.2878 and 0.2213, corresponding to relative improvements of 29.0% and 34.5%. Using the same backbone and descriptor set, Concat-Linear obtains QMOF and HMOF Mean MAEs of 0.5072 and 0.2573, while Concat-FFN improves them to 0.4760 and 0.2262. Cross-attention therefore reduces MAE by 43.3% and 14.0% relative to Concat-Linear, and by 39.5% and 2.2% relative to Concat-FFN, on QMOF and HMOF Mean, respectively. We also observe that cross-attention converges faster than Concat-FNN fusion on both datasets. A more detailed convergence analysis is provided in Appendix D.2.

4.4 Property-Specific Descriptor Complementarity

RQ4: Which descriptor families are most useful for each property?

The descriptor ablations in Table 2 reveal different requirements for QMOF and HMOF. Our multimodal model using the Bond Connectivity descriptor group achieves the lowest QMOF MAE among all our descriptor combinations, outperforming Metal–Ligand Geometry at 0.2847, Metal Coordination at 0.2883, Pore Structure at 0.2970, and All Descriptors at 0.2878. Adding every descriptor family therefore does not improve band-gap prediction over the compact bonding representation.

In contrast, the descriptor ablation results indicate that HMOF is primarily driven by pore-related information. Pore Structure alone obtains an HMOF Mean MAE of 0.2221, compared with 0.2411 for Bond Connectivity and approximately 0.243 for the metal-centered descriptor families. Combining Bonding and Pore produces the numerically lowest mean of 0.2209. Within the Pore Structure descriptor group, we further test whether its performance is driven by the compact AV + ASA subset. AV + ASA obtains an HMOF Mean MAE of 0.2229, while Bonding + AV + ASA obtains 0.2213.

Appendix Figure 5 provides a complementary qualitative view of the learned property-specific interactions. For QMOF, attention concentrates around metal-node and bonding-related MOFid com-

Table 3: Test MAE with 500 and 2K labeled examples, averaged over three seeds.

Backbone	Configuration	QMOF	HMOF: CH_4			HMOF: CO_2			HMOF
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5	Mean
500 labeled examples									
PMTransformer	Pretrained	0.5284	0.0347	0.1663	0.4332	0.1930	0.6282	1.2343	0.4483
Granite	All Descriptors	0.4807	0.0432	0.1643	0.3917	0.1776	0.5266	0.9835	0.3811
Granite + PubChem + Article	All Descriptors	0.4517	0.0401	0.1493	0.3539	0.1677	0.4803	0.8642	0.3426
2K labeled examples									
PMTransformer	Pretrained	0.4322	0.0239	0.1217	0.2967	0.1579	0.4757	0.9506	0.3378
Granite	All Descriptors	0.3647	0.0345	0.1422	0.3333	0.1543	0.4544	0.7886	0.3179
Granite + PubChem + Article	All Descriptors	0.3458	0.0321	0.1328	0.3109	0.1482	0.4272	0.7316	0.2971

ponents, consistent with the Bond Connectivity ablation, whereas HMOF shows broader interactions between pore descriptors and framework-level components, including the catenation label. To probe these associations quantitatively, we conduct two modality-corruption experiments by replacing either the full MOFid or only its catenation label with the corresponding content from another randomly selected test example. Shuffling the full MOFid substantially degrades performance on QMOF and all HMOF tasks, confirming the importance of correctly aligned MOFid information.

Restoring the correctly aligned MOFid while keeping the catenation label shuffled recovers the original QMOF performance and improves all HMOF results relative to full MOFid shuffling. However, HMOF performance remains below the fully unshuffled setting, with HMOF Mean increasing from 0.2213 to 0.5627, while QMOF remains nearly unchanged (0.2878 to 0.2884). Together, the corruption experiments and attention patterns in Appendix Figure 5 provide consistent evidence that the catenation label plays a key role in connecting MOFid representations with pore-structure descriptors for HMOF adsorption prediction.

4.5 Data-Efficient Transfer

RQ5: Does our multimodal model transfer effectively under limited supervision?

As shown in Table 12, in the setting with 500 labeled examples, continued pretraining improves the multimodal model, decreasing QMOF MAE from 0.4807 to 0.4517 and HMOF Mean from 0.3811 to 0.3426. Overall, the domain-adapted multimodal model performs best on QMOF, two of the three CH_4 targets, all three CO_2 targets, and HMOF Mean, while PMTransformer performs best on the CH_4 target at 0.05. These results demonstrate effective transfer of domain-adapted MOFid representations under limited supervision. With the larger 2K-label training set, continued pretraining similarly improves the multimodal model, decreasing QMOF MAE from 0.3647 to 0.3458 and HMOF Mean from 0.3179 to 0.2971. Overall, PMTransformer performs best on all three CH_4 targets, while our multimodal model performs best on QMOF, all three CO_2 targets, and HMOF Mean.

5 Conclusion

We introduced a two-stage framework for MOF property prediction that combines domain-adaptive continued pretraining with lightweight structural grounding. By adapting Granite to chemical literature and integrating frozen MOFid representations with interpretable CIF-derived descriptors through cross-attention, the framework captures complementary textual and structural information. Continued pretraining provides the largest gains for QMOF band-gap prediction, while pore and bonding descriptors drive improvements in HMOF adsorption. The proposed model achieves the lowest MAE across all three CO_2 targets and the HMOF Mean, with strong performance even under limited supervision. These results demonstrate that literature adaptation and compact physical grounding offer complementary, property-dependent benefits for MOF prediction.

AI USE STATEMENT

The authors did not use generative AI tools in the preparation of this work.

ETHICS STATEMENT

This study uses publicly available datasets and does not involve human participants or personal data. Continued pretraining uses MOF-related abstracts and articles from the PubMed Central Open Access Subset only.

REPRODUCIBILITY STATEMENT

The main paper describes the QMOF and HMOF datasets, adopted splits, model architecture, prediction settings, baselines, and evaluation protocol. The appendix provides definitions and preprocessing procedures for the 18 CIF-derived descriptors, continued-pretraining and downstream-training hyperparameters, model and parameter details, baseline implementation protocols, and matched settings for the cross-attention and concatenation comparison. It also documents the cross-attention analysis and individual results with means and standard deviations for all random seeds.

REFERENCES

PMC Open Access Subset. https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

Alauddin Ahmed, Saona Seth, Justin Purewal, Antek Wong-Foy, Michael Veenstra, Adam Matzger, and Donald Siegel. Exceptional hydrogen storage achieved by screening nearly half a million metal-organic frameworks. *Nature Communications*, 10, 04 2019. doi: 10.1038/s41467-019-09365-w.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013. doi: 10.1103/PhysRevB.87.184115. URL https://link.aps.org/doi/10.1103/PhysRevB.87.184115.

Anastasiya Bavykina, Nikita Kolobov, Il Son Khan, Jeremy Bau, Adrian Ramirez, and Jorge Gascon. Metal–organic frameworks in heterogeneous catalysis: Recent progress, new trends, and future perspectives. *Chemical Reviews*, 120, 03 2020. doi: 10.1021/acs.chemrev.9b00685.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371/.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031/.

Peter Boyd, Arunraj Chidambaram, Enrique García Diez, Christopher Ireland, Thomas Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Moosavi, Mercedes Maroto-Valer, Jeffrey Reimer, Jorge Navarro, Tom Woo, Susana García, Kyriakos Stylianou, and Berend Smit. Data-driven design of metal–organic frameworks for wet flue gas co2 capture. *Nature*, 576:253–256, 12 2019. doi: 10.1038/s41586-019-1798-7.

Benjamin J. Bucior, Andrew S. Rosen, Maciej Haranczyk, Zhenpeng Yao, Michael E. Ziebel, Omar K. Farha, Joseph T. Hupp, J. Ilja Siepmann, Alán Aspuru-Guzik, and Randall Q. Snurr. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Crystal Growth & Design*, 19(11):6682–6697, 2019. doi: 10.1021/acs.cgd.9b01050. URL https://doi.org/10.1021/acs.cgd.9b01050.

Pieremanuele Canepa, Calvin A. Arter, Eliot M. Conwill, Daniel H. Johnson, Brian A. Shoemaker, Karim Z. Soliman, and Timo Thonhauser. High-throughput screening of small-molecule adsorption in mof. *J. Mater. Chem. A*, 1:13597–13604, 2013. doi: 10.1039/C3TA12395B. URL http://dx.doi.org/10.1039/C3TA12395B.

Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: Self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023a. doi: 10.1021/jacs.2c11420. URL https://doi.org/10.1021/jacs.2c11420. PMID: 36706365.

Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: Self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023b. doi: 10.1021/jacs.2c11420. URL https://doi.org/10.1021/jacs.2c11420. PMID: 36706365.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Beatriz Cordero, Verónica Gómez, Ana E. Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. *Dalton Transactions*, 37(21):2832–2838, 01 2008. ISSN 1477-9226. doi: 10.1039/b801115j. URL https://doi.org/10.1039/b801115j.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Michael Fernandez, Peter G. Boyd, Thomas D. Daff, Mohammad Zein Aghaji, and Tom K. Woo. Rapid and accurate machine learning recognition of high performing metal organic frameworks for co2 capture. *The Journal of Physical Chemistry Letters*, 5(17):3056–3060, 09 2014. ISSN 1948-7185. doi: 10.1021/jz501331m. URL https://doi.org/10.1021/jz501331m.

Granite Team, IBM. Granite 4.0 350m. https://huggingface.co/ibm-granite/granite-4.0-350m, October 2025. Hugging Face model card, released October 28, 2025.

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, May 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00784-w. URL https://www.nature.com/articles/s41524-022-00784-w.

IBM Granite Team. Granite 3.3 2b model card. https://huggingface.co/ibm-granite/granite-3.3-2b-base, 2025. Accessed: 2026-09-25.

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

Yeonghun Kang, Hyunsoo Park, Berend Smit, and Jihan Kim. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nature Machine Intelligence*, 5(3):309–318, 2023.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1059. URL https://doi.org/10.1093/nar/gkae1059.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather J Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature communications*, 11(1):4068, 2020a.

Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G. Boyd, Yongjin Lee, Berend Smit, and Heather J. Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature Communications*, 11(1):4068, August 2020b. ISSN 2041-1723. doi: 10.1038/s41467-020-17755-8. URL https://www.nature.com/articles/s41467-020-17755-8%C2%BB.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2012.10.028. URL https://www.sciencedirect.com/science/article/pii/S0927025612006295.

Michael O'Keeffe, Maxim A. Peskov, Stuart J. Ramsden, and Omar M. Yaghi. The reticular chemistry structure resource (rcsr) database of, and symbols for, crystal nets. *Accounts of Chemical Research*, 41(12):1782–1789, 2008. doi: 10.1021/ar800124u. URL https://doi.org/10.1021/ar800124u. PMID: 18834152.

Hillary Pan, Alex M. Ganose, Matthew Horton, Muratahan Aykol, Kristin A. Persson, Nils E. R. Zimmermann, and Anubhav Jain. Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic Chemistry*, 60(3):1590–1603, 2021. doi: 10.1021/acs.inorgchem.0c02996. URL https://doi.org/10.1021/acs.inorgchem.0c02996. PMID: 33417450.

Hyunsoo Park, Yeonghun Kang, and Jihan Kim. Enhancing structure–property relationships in porous materials through transfer learning and cross-material few-shot learning. *ACS Applied Materials Interfaces*, 15(48):56375–56385, 12 2023a. ISSN 1944-8244. doi: 10.1021/acsami.3c10323. URL https://doi.org/10.1021/acsami.3c10323.

Hyunsoo Park, Yeonghun Kang, and Jihan Kim. Enhancing structure–property relationships in porous materials through transfer learning and cross-material few-shot learning. *ACS Applied Materials Interfaces*, 15(48):56375–56385, 12 2023b. ISSN 1944-8244. doi: 10.1021/acsami.3c10323. URL https://doi.org/10.1021/acsami.3c10323.

Andrew Rosen, Shaelyn Iyer, Debmalya Ray, Zhenpeng Yao, Alán Aspuru-Guzik, Laura Gagliardi, Justin Notestein, and Randall Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4, 04 2021a. doi: 10.1016/j.matt.2021.02.015.

Andrew Rosen, Victor Fung, Patrick Huck, Cody O'Donnell, Matthew Horton, Donald Truhlar, Kristin Persson, Justin Notestein, and Randall Snurr. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *npj Computational Materials*, 8:112, 05 2022. doi: 10.1038/s41524-022-00796-6.

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, 2021b.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

David Weininger. 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, 1988. doi: 10.1021/ci00057a005. URL `https://doi.org/10.1021/ci00057a005`.

Christopher Wilmer, Michael Leaf, Chang Yeon Lee, Omar Farha, Brad Hauser, Joseph Hupp, and Randall Snurr. Large-scale screening of hypothetical metal-organic frameworks. *Nature chemistry*, 4:83–9, 02 2012. doi: 10.1038/nchem.1192.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018a. doi: 10.1103/PhysRevLett.120.145301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.120.145301`.

Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018b.

Rui-Zhi Zhang, Sohan Seth, and James Cumby. Grouped representation of interatomic distances as a similarity measure for crystal structures. *Digital Discovery*, 2(1):81–90, 02 2023. ISSN 2635-098X. doi: 10.1039/d2dd00054g. URL `https://doi.org/10.1039/d2dd00054g`.

Hong-Cai Zhou, Jeffrey Long, and Omar Yaghi. Introduction to metal-organic frameworks. *Chemical reviews*, 112:673–4, 02 2012. doi: 10.1021/cr300014x.

MOF Literature Articles and Abstracts

O--H···O hydrogen bonds involving the coordinated and uncoordinated water molecules and intralayer π--π interactions between the pyridine rings [centroid--centroid distances = 3.571(2) and 3.569(2) Å] are observed.
Each layer interacts with two neighboring layers via Ag···O(H_2O) contacts and interlayer π--π interactions [centroid--centroid distances = 3.479(3) to 3.530(3) Å], leading to a three-dimensional supramolecular network.
The naphthalene-1,4-dicarboxylate and 1,1′-(butane-1,4-diyl)diimidazole ligands bridge the Ni centres to form a two-dimensional $(4,4)$ network, and O--H···O hydrogen bonds complete the structure.
The carboxylate group bridges adjacent Cu atoms, forming an infinite zigzag chain extending parallel to [001].
O--H···O hydrogen bonds between water molecules and carboxylate O atoms help consolidate the crystal packing.

Structural Dataset (PubChem)

Compound_Name : PE-Cer 12:1;2O/15:0
IUPAC_Name : 2-aminoethyl [(E)-3-hydroxy-2-(pentadecanoylamino)dodec-4-enyl] hydrogen phosphate
InChI : InChI=1S/C29H59N2O6P/c1-3-5-7-9-11-12-13-14-15-17-19-21-23-29... ...
InChIKey : NSYKRAGKTMQFFE-LSDHQDQOSA-N
Canonical_SMILES : CCCCCCCCCCCCCCC(=O)NC(COP(=O)(O)OCCN)C(C=CCCCCCCC)O
Isomeric_SMILES : CCCCCCCCCCCCCCC(=O)NC(COP(=O)(O)OCCN)C(/C=C/CCCCCCC)O
Molecular_Formula : C29H59N2O6P
Molecular_Weight : 562.8 g/mol
...

Figure 2: Examples of MOF articles and the PubChem structural dataset from our pre-training corpora.

APPENDIX

A CONSTRUCTION OF THE PRE-TRAINING MOF CORPORA

We construct the pre-training corpus from two complementary data sources: the PubMed Central (PMC) Open Access Subset and PubChem. The PMC Open Access Subset contains 3.4 million journal articles and preprints distributed under permissive licenses. To identify literature relevant to metal-organic frameworks (MOFs), we retrieve documents using domain-specific keywords, including *MOF*, *metal-organic framework*, *pore size*, *CO_2 adsorption*, *CO_2/N_2 selectivity*, *CO_2/H_2O selectivity*, *CO_2/CH_4 selectivity*, *CH_4 adsorption*, and *CO_2 uptake*. This filtering process yields 28,525 MOF-related articles and abstracts. We then apply the NLTK sentence tokenizer to segment each document into individual sentences, with each sentence stored on a separate line for continual pre-training.

To complement the scientific literature with structured chemical information, we incorporate records from PubChem, a large repository of publicly accessible chemical compounds. Each PubChem record contains information such as the compound name, IUPAC name, InChI, InChIKey, SMILES representations, molecular formula, molecular weight, and other physicochemical properties. We format the selected properties of each compound as a single textual record, allowing the model to learn the relationships among chemical identifiers, molecular structures, formulas, and properties. Combining MOF-related literature with structured PubChem records provides the model with both domain-specific scientific language and structured chemical knowledge. Figure 2 presents representative examples from the two data sources.

B Assessment of PubChem Pretraining Overlap with hMOF/QMOF

To assess possible pretraining data leakage, we performed an exact text-overlap analysis between the MOFid representations used as inputs to the LLM component and the PubChem records used for pretraining. This check is appropriate for our setting because the LLM branch of the multimodal model consumes MOFid as text and uses this representation to capture chemical semantics. For each MOF entry, we removed the topology suffix after `&&` from the MOFid and split the remaining chemical string by (.) to obtain individual MOFid components. These components were then compared by exact string matching against the PubChem SMILES strings used during pretraining, including Canonical SMILES and Isomeric SMILES fields when available. No chemical standardization, neutralization, or graph-based equivalence matching was applied, since the objective was to detect exact text-level exposure of the model input during LLM pretraining.

Across all HMOF and QMOF train, validation, and test splits, no exact overlaps were found between the extracted MOFid components and the PubChem SMILES strings in the sampled pretraining corpus. This outcome is reasonable given that the PubChem subset used for pretraining contained 250K sampled entries, which represents only a small fraction of the full PubChem compound space, reported to contain over 119 million compounds in the PubChem 2025 update. Therefore, even if some MOF-related components exist somewhere in the full PubChem database, the probability of observing an exact text match in a random 250K entry subset is low unless such overlaps are frequent. These results indicate that, in the text representation (MOFid) used to pre-train our LLM component, we found no evidence of exact SMILES-level leakage from the sampled PubChem pretraining corpus into the HMOF or QMOF downstream datasets.

C Structural Descriptor Computation

The following four families of descriptors are extracted directly from CIF files using CrystalNN-based neighbor detection (Pan et al., 2021).

Metal–Ligand Geometry. For every metal site in the structure, CrystalNN identifies the nearest-neighbor atoms. The Euclidean distance between each metal site and each of its neighbors is computed and collected across all metal sites in the unit cell. The minimum, mean, and maximum of these distances are retained as three separate descriptors.

Metal Coordination. For each metal site, the coordination number is defined as the number of neighbors identified by CrystalNN. The minimum, maximum, arithmetic mean, and population standard deviation of the coordination numbers across all metal sites in the unit cell are retained as four separate descriptors.

Bond Connectivity. A bond graph is constructed over the full unit cell using the CrystalNN local-environment strategy. Each bond is classified by the element types of its two endpoint atoms. Five bond-type counts are accumulated: carbon–carbon, carbon–oxygen, carbon–nitrogen, metal–oxygen, and metal–nitrogen. Each count reflects the total number of such bonds present in the unit cell, spanning both the organic linkers and the metal nodes.

Pore Structure. We compute six lightweight geometric proxies from the primitive unit cell, falling back to the original cell when necessary. Consistent with structural representations using covalent radii and interatomic distances (Cordero et al., 2008; Zhang et al., 2023), atoms are modelled as hard spheres with approximate covalent radii and a default radius of 1.5 Å for unlisted elements. Void fraction is estimated from the cell volume after subtracting the summed atomic-sphere volumes. Accessible volume is derived from the estimated void fraction and cell volume, while accessible surface area is approximated from the unit-cell geometry. Pairwise-distance statistics define the largest-cavity and pore-limiting proxies, while the largest-free-sphere proxy is derived from the largest-cavity value. These descriptors serve as lightweight structural conditioning features rather than rigorous probe-accessible pore properties.

D Fusion Architecture Details

D.1 Architecture Design

Both the cross-attention model and the concatenation ablation share the same LLM backbone (Granite-3.3-2B, frozen), the same 18 structural descriptors, the same `large` preset hyperparameters (`fusion_dim`=1024, `num_layers`=6, `ffn_dim`=4096), and the same regression head. We compare two concatenation variants: Concat-Linear uses a single projection layer after merging the two modalities, while Concat-FFN applies 6 FFN layers of identical depth and width to the cross-attention model. The sole difference between Concat-FFN and the cross-attention model is the presence of the multi-head attention sub-layer inside each block. For the concatenation variants, the LLM sequence is reduced to a single vector via mean-pooling rather than retaining the full token sequence; concatenating the full sequence with a global descriptor vector would require broadcasting the descriptor across all token positions, introducing implicit token-level interactions through the subsequent FFN layers that would confound the ablation—mean-pooling is the parameter-free, uniform-weight baseline that stands in direct contrast to the descriptor-guided, learned aggregation performed by cross-attention.

Table 4 lists the components shared by all three models. Table 5 lists the components that differ.

Table 4: Shared components across all three models.

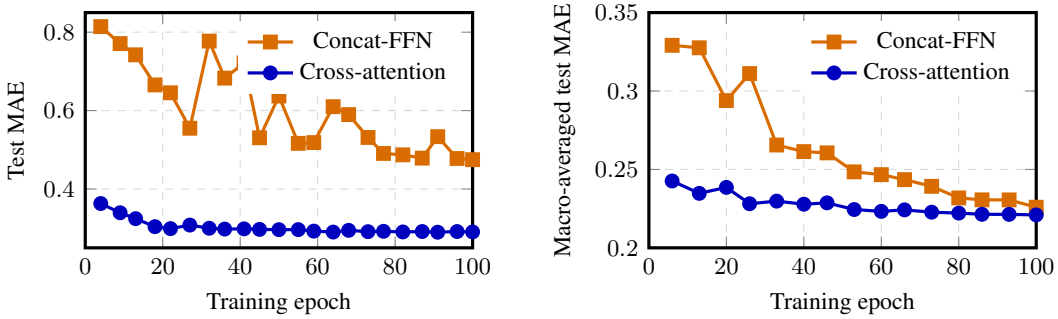
Component	Value
LLM	Granite-3.3-2B (frozen)
Descriptors	18 features
Fusion dim	1024
FFN width	4096
Head	$1024 \rightarrow 512 \rightarrow 128 \rightarrow$ out
Loss	MSE

Table 5: Fusion mechanism differences.

	Cross-Attention	Concat-FFN	Concat-Linear
LLM tokens	Full sequence	Mean-pooled	Mean-pooled
Merging	Q attends K, V	Concat $\rightarrow$ Linear	Concat $\rightarrow$ Linear
Attention	16 multi-head attention	—	—
FFN blocks	6	6	—
Merge layer	—	—	Linear + GELU + LN

D.2 Convergence of the Fusion Strategies

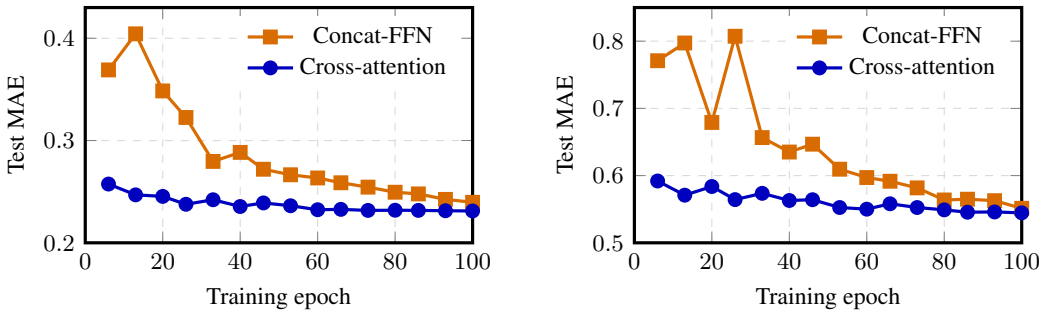
Figures 3 and 4 reveal a consistent difference in the convergence behavior of the two fusion strategies. Cross-attention converges rapidly, reaching a low-MAE regime within the early training epochs and subsequently exhibiting only minor variations across checkpoints. This behavior is especially pronounced on QMOF, where the test MAE stabilizes after approximately 20 epochs, while Concat-FFN converges more slowly and continues to fluctuate substantially throughout training. A similar pattern is observed on HMOF: Cross-attention achieves a relatively stable macro-averaged test MAE early in training, whereas Concat-FFN follows a more gradual and less stable trajectory. The target-level curves further show that the slower convergence of Concat-FFN is particularly evident for the high-pressure CH_4 and CO_2 tasks, with pronounced early fluctuations followed by a gradual decline in test MAE. In contrast, Cross-attention maintains smoother trajectories and reaches low test MAE values within fewer epochs across both targets. These results suggest that explicitly modeling interactions between modalities enables more effective and stable information integration. Test-set trajectories were examined only after training to compare convergence behavior and were not used for model or checkpoint selection.



(a) QMOF test MAE.

(b) HMOF macro-averaged test MAE.

Figure 3: Test MAE at matched training checkpoints for Cross-attention and Concat-FFN on QMOF and HMOF. The HMOF value is the unweighted macro-average over the six adsorption targets. Test-set evaluations were used only for comparison and not for model or checkpoint selection.



(a) CH_4 adsorption at 2.5 bar.

(b) CO_2 adsorption at 2.5 bar.

Figure 4: Test MAE at matched training checkpoints for the high-pressure HMOF adsorption targets.

E Hyperparameter Choice

E.1 LLM Hyperparameters

The hyperparameter setup for pre-training is detailed in Table 6. Our experimental setup uses a single H100 80GB GPU for pre-training the LLM and for fine-tuning, where we use the Transformers (Wolf et al., 2020), Transformer Reinforcement Learning (TRL) (von Werra et al., 2020), and FlashAttention-2 (Dao et al., 2022; Dao, 2024) libraries. During pre-training, although we maintained a batch size of 1, we extended the maximum sequence length to 4096. This was achieved with the packing feature enabled, a choice made to ensure sequences are densely packed rather than padded.

Table 6: Hyperparameter configurations for pre-training and LLM-based MOF property prediction. Packing refers to packing multiple sequences into a single batch to improve training efficiency.

Hyperparameter	Pre-Training	MOF Prediction
Training Configuration		
Epochs	1	5
Batch Size	1	16
Learning Rate	5e-5	3e-5
Packing	True	False
Gradient Checkpointing	True	True
BF16 Precision	True	True
Sequence Configuration		
Max Sequence Length	4096	512
Max New Tokens	Unsupervised	256

E.2 Multimodal Hyperparameters

The hyperparameter settings for our multimodal models are detailed in Table 7. These configurations were optimized on the validation set and applied to train all our models. All multimodels were trained using a single H100 80GB GPU.

Table 7: Hyperparameters for MultiModal MOF Property Prediction.

Hyperparameter	Value
Training Configuration	
Epochs	100
Train Batch Size	24
Warmup Steps	100
Number of Workers (CIF Processing)	80
BF16	True
Max Text Length	256
Fusion Adapter (Large)	
Fusion Dimension	1024
Number of Attention Heads	16
Number of Layers	6
FFN Dimension	4096
Attention Dropout	0.0
FFN Dropout	0.1
Residual Dropout	0.0
Pre-normalization	True
Activation Function	GELU
Initialization Scale	1.0
LayerNorm Epsilon	1×10^{-5}
Regression Head (Large)	
Input Dimension	1024
Hidden Dimensions	[512, 128]
Activation Function	GELU
Dropout	0.05

F Impact of Model Size in MultiModal

We study the impact of model scale using five predefined multimodal configurations spanning a wide capacity range, as illustrated in Table 8.

The regression head is scaled consistently with the fusion dimension for each configuration. All other architectural parameters are held constant across model sizes, including GELU activations,

Table 8: Fusion model configurations across different scales. Parameter counts are reported for each model size.

Model Size	Params	Fusion Dim.	Attn. Heads	Fusion Layers	FFN Width
Medium	14.08M	512	8	4	2048
Large	78.80M	1024	16	6	4096
xLarge	410.50M	2048	32	8	8192
xxLarge	1.15B	3072	48	10	12288
xxxLarge	2.44B	4096	64	12	16384

pre-norm transformers, attention, and residual dropout set to zero, FFN dropout of 0.1, initialization scale 1.0, and layer-normalization epsilon of 10^{-5}.

Table 9: Impact of multimodal model size on MOF property prediction performance, evaluated using mean absolute error (MAE); lower values indicate better performance. Parameter counts are reported for each model.

Model	Params	QMOF	HMOF CH_4			HMOF CO_2			HMOF Mean
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5	
External structure-pretrained models									
MOFTransformer	85M	0.368	0.018	0.098	0.241	0.113	0.360	0.671	0.250
PMTransformer	85M	0.269	0.017	0.090	0.225	0.110	0.339	0.641	0.237
Our multimodal models									
Medium	14.08M	0.302	0.020	0.101	0.237	0.111	0.333	0.564	0.228
Large	78.80M	0.288	0.019	0.098	0.232	0.109	0.325	0.545	0.221
xLarge	410.50M	0.280	0.018	0.097	0.229	0.107	0.325	0.544	0.220
xxLarge	1.15B	0.279	0.019	0.097	0.227	0.106	0.322	0.541	0.219
xxxLarge	2.44B	0.282	0.019	0.098	0.229	0.108	0.323	0.541	0.219

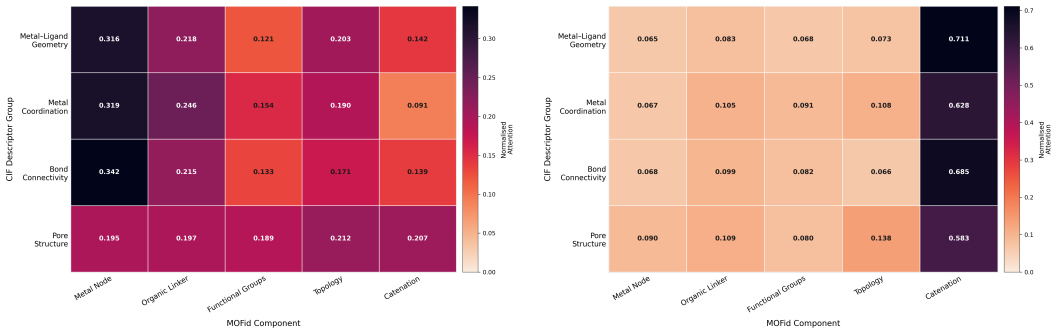
As shown in Table 9, increasing the multimodel size leads to clear performance improvements on QMOF, with MAE decreasing from 0.302 for the Medium model to 0.289 for Large and further to 0.280 for xLarge. This trend highlights that scaling the multimodel is particularly effective in low-resource settings (5.2K samples), where the model benefits from richer utilization of CIF-derived structural signals. In contrast, gains on HMOF are more modest, with average MAE improving gradually from 0.228 to 0.219 as the model size increases. This suggests that when a large number of training samples (72K) is combined with full CIF structural information, the availability of strong supervision can partially offset the need for further increases in model size. Notably, improvements begin to saturate beyond the Large configuration, as xxLarge and xxxLarge variants provide only marginal or inconsistent gains, indicating that the Large multimodel achieves the best balance between predictive performance and computational cost.

Based on these observations, we choose to conduct subsequent experiments with the **Large** configuration, as it represents the most effective trade-off between accuracy gains and computational efficiency. While larger variants slightly improve performance in some cases, their benefits diminish relative to the substantial increase in model capacity and training cost. The **Large** multi-model therefore provides a robust and scalable setting that captures the advantages of increased representational power without incurring unnecessary complexity.

F.1 Task-Specific Cross-Attention Analysis

The descriptor-family and combination ablations reported in Table 2 show that Bond Connectivity provides the strongest individual descriptor configuration for QMOF, while Pore Structure provides the strongest individual configuration for HMOF. However, the numerical ablations do not indicate which parts of the MOFid representation benefit most from interaction with these descriptor families. Examining these interactions provides additional evidence for our main research question concerning the complementary roles of MOFid representations and structural descriptors.

To address this question, Figure 5 presents cross-attention heat maps showing how each CIF-derived descriptor family attends to five chemically meaningful MOFid components: Metal Node, Organic Linker, Functional Groups, Topology, and Catenation. Token-level attention scores are mapped to



(a) QMOF **(b) HMOF**

Figure 5: Consolidated single-family cross-attention heat maps for QMOF and HMOF. Each row is obtained from a model trained using only the corresponding descriptor family and is normalized independently across the MOFid components. Values represent learned attention patterns rather than causal feature importance.

these components using the tokenizer's character-offset mappings. Scores are averaged across the tokens within each component, normalized across the five components so that each row sums to one, and then averaged over the test set.

To isolate the associations learned by each descriptor family, each heat-map row is obtained from a model trained using only the corresponding descriptor family. The four independently obtained rows are then stacked to form a consolidated 4×5 matrix. Because the learned associations differ substantially between QMOF and HMOF, the two tasks are analyzed separately.

For HMOF, Pore Structure attends most strongly to the Catenation component of MOFid. Combined with the comparatively strong HMOF performance obtained using Pore Structure in Table 2, this pattern suggests that pore-related descriptors complement the framework-level structural context encoded by Catenation in MOFid.

Catenation describes the degree of framework interpenetration, in which multiple framework nets occupy the same crystal volume. Such interpenetration can alter accessible pore space, surface area, pore dimensions, and adsorption-site accessibility. The strong relative attention between Pore Structure and Catenation is therefore consistent with the dependence of gas adsorption on global framework geometry. However, this pattern represents a learned association and does not establish that the improved HMOF performance is caused by the Catenation component.

For QMOF, Metal–Ligand Geometry, Metal Coordination, and Bond Connectivity attend most strongly to the Metal Node. This pattern is chemically plausible because QMOF band-gap prediction depends on the electronic structure of the framework, which can be influenced by metal identity, coordination environment, and bonding interactions between the metal centers and organic linkers. In particular, the strong Metal Node association of Bond Connectivity complements its favorable numerical performance in the descriptor ablation. This pattern suggests that explicit connectivity information complements the metal-related chemical context encoded in MOFid during band-gap prediction.

Overall, the heat-map analysis shows task-specific associations between the descriptor families and MOFid components. These attention profiles complement the numerical ablations presented in Table 2.

G Full Results

Table 10: Full-data test MAE on QMOF and six HMOF adsorption tasks. Values are reported as mean ± sample standard deviation across three seeds.

Backbone	Configuration	QMOF	HMOF: CH_4			HMOF: CO_2		
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5
External baselines								
Random Forest	CIF Descriptors	0.5117 ± 0.0010	0.0248 ± 0.0001	0.1386 ± 0.0001	0.3439 ± 0.0004	0.1446 ± 0.0000	0.4492 ± 0.0003	0.7667 ± 0.0006
XGBoost	CIF Descriptors	0.5181 ± 0.0032	0.0278 ± 0.0002	0.1543 ± 0.0003	0.3875 ± 0.0006	0.1570 ± 0.0005	0.4962 ± 0.0013	0.8607 ± 0.0032
SOAP	Atomic Environment	0.4240 ± 0.0070	0.0220 ± 0.0010	0.1060 ± 0.0010	0.2390 ± 0.0020	0.1150 ± 0.0020	0.3390 ± 0.0040	0.6660 ± 0.0030
CGCNN	Scratch	0.2750 ± 0.0150	0.0280 ± 0.0010	0.1210 ± 0.0060	0.3330 ± 0.0170	0.1260 ± 0.0050	0.3910 ± 0.0170	0.8180 ± 0.0500
CGCNN (400K MOF Structures)	Pretrained	**0.2560 ± 0.0060**	0.0250 ± 0.0010	0.0990 ± 0.0010	0.2580 ± 0.0080	0.1100 ± 0.0010	0.3300 ± 0.0020	0.6450 ± 0.0030
MOFormer	Scratch	0.3870 ± 0.0010	0.0340 ± 0.0000	0.1740 ± 0.0020	0.3850 ± 0.0030	0.1780 ± 0.0020	0.5580 ± 0.0010	1.0000 ± 0.0130
MOFormer (400K MOFid–Structure)	Pretrained	0.3670 ± 0.0050	0.0330 ± 0.0000	0.1610 ± 0.0110	0.3840 ± 0.0030	0.1580 ± 0.0010	0.5450 ± 0.0080	0.9820 ± 0.0110
MOFTransformer (1.1M MOF Structures)	Pretrained	0.3905 ± 0.0090	0.0181 ± 0.0005	0.0977 ± 0.0023	0.2409 ± 0.0083	0.1132 ± 0.0008	0.3601 ± 0.0044	0.6710 ± 0.0066
PMTransformer (1.9M Porous Structures)	Pretrained	0.2692 ± 0.0014	**0.0174 ± 0.0003**	**0.0896 ± 0.0005**	**0.2248 ± 0.0010**	0.1103 ± 0.0014	0.3394 ± 0.0046	0.6406 ± 0.0097
Our LLM models								
Granite 3.3-2B	Regression	0.4054 ± 0.0025	0.0313 ± 0.0006	0.1485 ± 0.0007	0.3433 ± 0.0007	0.1718 ± 0.0002	0.4873 ± 0.0010	0.8445 ± 0.0027
Granite 3.3-2B	Generative	0.3513 ± 0.0131	0.0193 ± 0.0004	0.1026 ± 0.0014	0.2481 ± 0.0029	0.1220 ± 0.0012	0.3724 ± 0.0048	0.6649 ± 0.0080
Granite 3.3-2B + PubChem	Generative	0.3334 ± 0.0115	0.0191 ± 0.0003	0.1020 ± 0.0012	0.2467 ± 0.0026	0.1207 ± 0.0014	0.3676 ± 0.0039	0.6573 ± 0.0065
Granite 3.3-2B + PubChem + Articles	Generative	0.3130 ± 0.0024	0.0192 ± 0.0001	0.1023 ± 0.0004	0.2465 ± 0.0015	0.1201 ± 0.0000	0.3661 ± 0.0012	0.6504 ± 0.0029
Our multimodal models								
Granite 3.3-2B + PubChem + Articles	All Descriptors	0.2878 ± 0.0029	0.0190 ± 0.0001	0.0978 ± 0.0005	0.2318 ± 0.0011	0.1087 ± 0.0002	**0.3252 ± 0.0012**	0.5450 ± 0.0004
Granite 3.3-2B + PubChem + Articles	Bond Connectivity	0.2796 ± 0.0047	0.0202 ± 0.0001	0.1046 ± 0.0000	0.2466 ± 0.0003	0.1146 ± 0.0009	0.3488 ± 0.0020	0.6115 ± 0.0021
Granite 3.3-2B + PubChem + Articles	Pore Structure	0.2970 ± 0.0042	0.0192 ± 0.0003	0.0984 ± 0.0003	0.2319 ± 0.0007	0.1092 ± 0.0007	0.3262 ± 0.0015	0.5477 ± 0.0023
Granite 3.3-2B + PubChem + Articles	Bonding + Pore	0.2862 ± 0.0040	0.0189 ± 0.0003	0.0974 ± 0.0009	0.2299 ± 0.0014	**0.1083 ± 0.0011**	0.3260 ± 0.0026	**0.5447 ± 0.0016**
Backbone and scale comparisons without CPT								
Granite 4.0-350M	All Descriptors	0.3364 ± 0.0036	0.0202 ± 0.0000	0.1044 ± 0.0005	0.2475 ± 0.0005	0.1147 ± 0.0016	0.3453 ± 0.0005	0.5847 ± 0.0004
Granite 3.3-2B	All Descriptors	0.3031 ± 0.0018	0.0197 ± 0.0002	0.1010 ± 0.0002	0.2379 ± 0.0009	0.1124 ± 0.0006	0.3336 ± 0.0011	0.5589 ± 0.0003
Phi-4 Mini 4B	All Descriptors	0.3122 ± 0.0023	0.0203 ± 0.0004	0.1039 ± 0.0005	0.2443 ± 0.0003	0.1141 ± 0.0010	0.3422 ± 0.0013	0.5778 ± 0.0041

Table 11: Full-data test MAE on QMOF and six HMOF adsorption tasks. Values are reported as mean ± sample standard deviation across three seeds.

Input Configuration	Fusion / Head	QMOF	HMOF: CH_4			HMOF: CO_2		
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5
Metal–Ligand	Cross-Attention	0.2847 ± 0.0038	0.0204 ± 0.0002	0.1055 ± 0.0004	0.2503 ± 0.0009	0.1152 ± 0.0006	0.3514 ± 0.0011	0.6169 ± 0.0034
Metal Coordination	Cross-Attention	0.2883 ± 0.0018	0.0203 ± 0.0002	0.1054 ± 0.0008	0.2498 ± 0.0010	0.1150 ± 0.0012	0.3523 ± 0.0017	0.6165 ± 0.0017
Bond Connectivity	Cross-Attention	**0.2796 ± 0.0047**	0.0202 ± 0.0001	0.1046 ± 0.0000	0.2466 ± 0.0003	0.1146 ± 0.0009	0.3488 ± 0.0020	0.6115 ± 0.0021
Pore Structure	Cross-Attention	0.2970 ± 0.0042	0.0192 ± 0.0003	0.0984 ± 0.0003	0.2319 ± 0.0007	0.1092 ± 0.0007	0.3262 ± 0.0015	0.5477 ± 0.0023
– AV + ASA	Cross-Attention	–	0.0191 ± 0.0003	0.0982 ± 0.0005	0.2331 ± 0.0007	0.1087 ± 0.0005	0.3267 ± 0.0013	0.5514 ± 0.0047
Metal–Ligand + Coordination	Cross-Attention	0.2831 ± 0.0019	0.0204 ± 0.0003	0.1061 ± 0.0003	0.2499 ± 0.0010	0.1155 ± 0.0004	0.3529 ± 0.0016	0.6205 ± 0.0026
Metal–Ligand + Coordination + Bonding	Cross-Attention	0.2803 ± 0.0012	0.0198 ± 0.0002	0.1021 ± 0.0005	0.2397 ± 0.0012	0.1130 ± 0.0003	0.3400 ± 0.0012	0.5882 ± 0.0019
Bonding + Pore	Cross-Attention	0.2862 ± 0.0040	0.0189 ± 0.0003	0.0974 ± 0.0009	**0.2299 ± 0.0014**	**0.1083 ± 0.0011**	0.3260 ± 0.0026	**0.5447 ± 0.0016**
– Bonding + AV + ASA	Cross-Attention	0.2858 ± 0.0025	**0.0188 ± 0.0002**	**0.0973 ± 0.0006**	0.2310 ± 0.0004	0.1085 ± 0.0009	0.3259 ± 0.0006	0.5464 ± 0.0021
All Descriptors	Concat-Linear	0.5072 ± 0.0031	0.0244 ± 0.0002	0.1185 ± 0.0004	0.2741 ± 0.0016	0.1299 ± 0.0003	0.3690 ± 0.0013	0.6279 ± 0.0003
All Descriptors	Concat-FFN	0.4760 ± 0.0019	0.0210 ± 0.0002	0.1027 ± 0.0002	0.2399 ± 0.0004	0.1122 ± 0.0006	0.3301 ± 0.0008	0.5512 ± 0.0003
All Descriptors	Cross-Attention	0.2878 ± 0.0029	0.0190 ± 0.0001	0.0978 ± 0.0005	0.2318 ± 0.0011	0.1087 ± 0.0002	**0.3252 ± 0.0012**	0.5450 ± 0.0004
– Full MOFid shuffle	Cross-Attention	1.0983 ± 0.0075	0.0619 ± 0.0015	0.3572 ± 0.0032	0.9147 ± 0.0030	0.3497 ± 0.0021	1.1814 ± 0.0022	2.3780 ± 0.0096
– MOFid Catenation shuffle	Cross-Attention	0.2884 ± 0.0017	0.0268 ± 0.0006	0.1564 ± 0.0046	0.4147 ± 0.0131	0.1579 ± 0.0008	0.5827 ± 0.0152	1.2464 ± 0.0350

Table 12: Test MAE with 500 and 2K labeled examples. Values are reported as mean ± standard deviation across three independent runs. Lower is better; bold and underlined values indicate the best and second-best means, respectively.

Backbone	Configuration	QMOF	HMOF: CH_4			HMOF: CO_2		
		Band gap	0.05	0.5	2.5	0.05	0.5	2.5
500 labeled examples								
PMTransformer	Pretrained	0.5284 ± 0.0097	**0.0347 ± 0.0032**	0.1663 ± 0.0039	0.4332 ± 0.0155	0.1930 ± 0.0063	0.6282 ± 0.0053	1.2343 ± 0.0165
Granite	All Descriptors	0.4807 ± 0.0213	0.0432 ± 0.0008	0.1643 ± 0.0046	0.3917 ± 0.0099	0.1776 ± 0.0048	0.5266 ± 0.0073	0.9835 ± 0.0444
Granite + PubChem + Article	All Descriptors	**0.4517 ± 0.0262**	0.0401 ± 0.0015	**0.1493 ± 0.0032**	**0.3539 ± 0.0099**	**0.1677 ± 0.0022**	**0.4803 ± 0.0032**	**0.8642 ± 0.0262**
2K labeled examples								
PMTransformer	Pretrained	0.4322 ± 0.0190	**0.0239 ± 0.0004**	**0.1217 ± 0.0023**	**0.2967 ± 0.0085**	0.1579 ± 0.0184	0.4757 ± 0.0619	0.9506 ± 0.1042
Granite	All Descriptors	0.3647 ± 0.0077	0.0345 ± 0.0008	0.1422 ± 0.0021	0.3333 ± 0.0041	0.1543 ± 0.0041	0.4544 ± 0.0066	0.7886 ± 0.0049
Granite + PubChem + Article	All Descriptors	**0.3458 ± 0.0039**	0.0321 ± 0.0004	0.1328 ± 0.0006	0.3109 ± 0.0017	**0.1482 ± 0.0017**	**0.4272 ± 0.0035**	**0.7316 ± 0.0060**