The Role of Geometric Representations in Coarse-Grained Graph Neural Networks for Metal–Organic Frameworks

Ahmed Arif, Dr. Bohui Lyu, Prof. Kangming Li

ahmed.arif@kaust.edu.sa, bohui.lyu@kaust.edu.sa, kangming.li@kaust.edu.sa

Abstract

Coarse-grained graph neural networks (GNNs) represent metal–organic frameworks (MOFs) as graphs of building blocks, offering substantial computational speedups for high-throughput screening while discarding explicit 3D void geometry. In this work, we present a systematic ablation of three progressively richer geometric representations in coarse-grained GNNs: pairwise metal–linker distances, line-graph bond angles, and global persistent-homology (PH) topological descriptors. Evaluating across nine electronic, structural, and multi-pressure gas adsorption targets on both experimental (QMOF) and computationally generated (hMOF) databases, we find that the predictive power of geometric features is governed primarily by dataset topological diversity. Global persistent-homology descriptors yield the strongest predictive signal, reducing adsorption and pore-size prediction errors by 53–73% over coarse-grained baselines and outperforming atomistic models. Pairwise distances and angles provide intermediate gains (20–47%), while geometric features become redundant when framework topology is strictly constrained by building-block chemistry. These findings establish that global pore-space topology is essential when geometry varies independently of composition, providing a principled framework for selecting geometric representations in reticular materials discovery.

Date: August 24, 2026

Code: https://github.com/ThugyBoy/mof-geometry-gnn

Contents

1 Introduction 3

2 Related Work 4

2.1 Graph Neural Networks for Crystalline Materials . . . 4
2.2 Coarse-Grained Representations for MOFs . . . 5
2.3 Geometric and Topological Descriptors for MOFs . . . 5
2.4 MOF Databases and Computational Screening . . . 6

3 Methodology 6

3.1 Datasets and Target Properties . . . 6
3.2 Coarse-Grained Graph Representation (CG2NN) . . . 7
3.3 Geometric Augmentations . . . 8
3.4 Persistent-Homology Topology Features . . . 8
3.5 Baseline Models . . . 9
3.6 Training Protocol . . . 9

4 Results 10

4.1 Overview . . . 10
4.2 QMOF: Geometry Is Neutral on Building-Block-Driven Properties . . . 10
4.3 hMOF: Geometry Strongly Improves Adsorption and PLD Prediction . . . 11
4.4 The Cross-Dataset PLD Confound . . . 12
4.5 Persistent Homology: The Strongest Geometry Signal on hMOF . . . 13
4.6 The Topology Reversal: hMOF vs. QMOF . . . 14
4.7 Atomistic GNNs: Local Cutoff vs. Global Pore Geometry . . . 14
4.8 Subgroup Analysis: Geometry Benefit Scales with Porosity . . . 15

5 Discussion 18

5.1 Why Geometry Helps: The Decoupling Principle . . . 18
5.2 The Length-Scale Dichotomy and Atomistic Model Failure . . . 18
5.3 Why Persistent Homology Outperforms Pairwise Geometry . . . 19
5.4 The Same Features, Different Models: Feature-Driven, Not Model-Driven . . . 19
5.5 Generalization and Boundaries . . . 20

6 Limitations & Future Work 20

6.1 Limitations . . . 20
6.2 Future Work . . . 21

7 Conclusion 23

References 23

1 Introduction

Metal–organic frameworks (MOFs) are a class of crystalline nanoporous materials composed of inorganic metal nodes connected by organic linkers. Their modular synthesis, tunable pore sizes, and enormous internal surface areas make them promising candidates for gas storage and separation applications, including carbon dioxide capture, methane storage, and hydrogen purification [1, 2]. The chemical and topological space of possible MOFs is vast: computational enumeration frameworks such as hMOF [1] have generated over 137,000 hypothetical structures, while databases like QMOF [2] provide density functional theory (DFT) properties for over 20,000 synthesized frameworks. Exhaustive experimental or computational screening of this space is infeasible, motivating the development of deep learning models that can predict MOF properties rapidly and accurately.

Graph neural networks (GNNs) have emerged as a dominant deep learning paradigm for crystalline materials. Atomistic GNNs such as CGCNN [3], MEGNet [4], and ALIGNN [5] represent a crystal as a graph of atoms connected by interatomic bonds, and learn property-relevant representations through message passing. These models achieve strong accuracy but incur significant computational cost on MOFs, whose unit cells often contain hundreds of atoms, making them impractical for high-throughput screening of large databases. Coarse-grained representations mitigate this cost by compressing a MOF into a graph of building blocks metal nodes and organic linkers rather than individual atoms. CG2NN [6] is one such coarse-grained approach that represents a MOF as a heterogeneous graph of metal and linker building blocks, reducing the graph size from hundreds of atom nodes to a few dozen building-block nodes. However, coarse-graining inevitably discards geometric information. In atomistic graphs, interatomic distances and bond angles are preserved in edge features; in coarse-grained graphs, these are replaced by metal–linker connectivity, and the explicit three-dimensional geometry of the framework pore sizes, channel networks, void topology is lost. This loss is potentially critical for adsorption properties, which are governed not by local bonding but by the geometry of the pore space: pore-limiting diameter, accessible surface area, and the topological structure of void channels and cavities all directly determine gas uptake [7]. Whether and how to reintroduce geometric information into coarse-grained MOF models remains an open question.

A natural hierarchy of geometric descriptors exists. At the simplest level, pairwise metal–linker distances can be encoded as radial basis function (RBF) edge features, analogous to the distance encoding used in atomistic GNNs. At a higher level, bond angles between building blocks can be captured via a line graph, mirroring the approach of ALIGNN. At the most comprehensive level, persistent homology (PH) [8] can capture the full topological structure of the pore space the distribution of channels (1-dimensional features) and cavities (2-dimensional features) across all length scales as a fixed-length vector. Krishnapriyan et al. [7] showed that such topological descriptors, combined with chemical embeddings in a random forest model, substantially outperformed traditional scalar pore descriptors (pore-limiting diameter, void fraction) for adsorption prediction. However, it remains unclear whether these findings transfer to graph neural network architectures, where the model can learn geometric relationships through message passing, and whether pairwise or angular geometry can substitute for the more expensive topological features.

In this work, we conduct a systematic ablation of geometric information in coarse-grained MOF models. Starting from a CG2NN baseline that encodes only building-block identity and connectivity, we progressively add three levels of geometric detail: (1) RBF-encoded metal–linker distances as edge features, (2) line-graph bond angles between adjacent building blocks, and (3) global persistent-homology descriptors of the pore space. We evaluate all variants on two

datasets with fundamentally different geometric characteristics: QMOF, comprising DFT-relaxed experimentally-synthesized MOFs where building-block identity largely determines framework geometry, and a reduced subset of the hMOF database comprising 51,163 computationally-generated hypothetical MOFs where linker lengths and metal–linker distances vary independently of chemistry. We compare against three atomistic baselines CGCNN, MEGNet, and ALIGNN and a classical random forest model reproducing the Krishnapriyan et al. pipeline.

Our key findings are threefold. First, the benefit of geometric features depends on the dataset, not the target: pairwise distances and angles are neutral on QMOF (where building-block identity fixes geometry) but reduce error by 20–47% on the reduced hMOF set (where continuum geometry varies independently of chemistry). Second, persistent-homology topology features are uniformly the strongest geometry signal on hMOF, outperforming pairwise and angular geometry on every target and beating all atomistic baselines by 53–73% relative to the coarse-grained baseline. Third, this advantage reverses on QMOF, where topology features are neutral to harmful confirming that the geometric signal is governed by dataset topological diversity rather than model capacity or target type. These results establish that the choice of geometric representation in coarse-grained MOF models should be guided by the geometric diversity of the target dataset, and that persistent-homology-based pore descriptors offer the most informative and computationally efficient geometry encoding for high-throughput screening of hypothetical MOF databases.

2 Related Work

2.1 Graph Neural Networks for Crystalline Materials

Graph neural networks have become the dominant deep learning paradigm for property prediction in crystalline solids. CGCNN [3] introduced the crystal graph representation, encoding a periodic structure as nodes (atoms) connected by edges to nearest neighbors within a cutoff distance, with Gaussian distance expansions as edge features and pooling over all atoms for the global readout. MEGNet [4] generalized this framework with set2set readout and a unified architecture for molecules and crystals, achieving strong performance on formation energies and band gaps across multiple datasets. ALIGNN [5] extended the crystal graph with an explicit line graph encoding interatomic bond angles, demonstrating that angular geometry improves prediction of electronic and thermodynamic properties on the JARVIS-DFT and Materials Project databases. These atomistic GNNs achieve state-of-the-art accuracy on properties governed by local atomic coordination band gaps, formation energies, elastic moduli but their receptive field is bounded by the cutoff distance (typically 6–12 Å), which becomes a limitation for properties governed by long-range structural order, as we demonstrate in this work for MOF adsorption.

Several recent efforts have applied atomistic GNNs specifically to MOFs. Choudhary et al. [9] benchmarked ALIGNN on the hMOF database (137,953 structures) for CO_2 adsorption, pore-limiting diameter (PLD), and largest cavity diameter (LCD). They reported reasonable adsorption accuracy (MAE ~0.48 mol/kg for CO_2 at 2.5 bar) despite poor PLD/LCD prediction, and concluded that "adsorption properties are largely determined by short-range interactions." Our results on a reduced hMOF subset contradict this conclusion: ALIGNN catastrophically fails on adsorption (+148% to +314% vs. coarse-grained baseline) where 41% of structures have PLD exceeding the 8 Å cutoff, and global topological features not local atomic interactions provide the decisive predictive signal. We attribute the discrepancy to differences in dataset composition and analyze the mechanism in Section 4.7. More recently, Uni-MOF [10] demonstrated that a unified transformer-based architecture pretrained on multiple MOF databases can achieve strong transfer learning for adsorption prediction, though it does not address the geometric representation

question that is the focus of this work.

2.2 Coarse-Grained Representations for MOFs

Coarse-graining compressing a MOF into a graph of building blocks rather than individual atoms offers a computationally efficient alternative to atomistic representations, particularly for high-throughput screening of large hypothetical databases. The building-block decomposition leverages the modular synthesis of MOFs: a framework is assembled from a finite set of inorganic secondary building units (SBUs) and organic linkers, and its net topology is determined by the connectivity of these components. CG2NN [6] represents a MOF as a heterogeneous graph with metal-node and linker-node types, using SAGE convolution with per-type readout. This reduces the graph from hundreds of atom nodes to a few dozen building-block nodes, enabling training on databases where atomistic GNNs are prohibitively expensive due to large unit cells.

However, coarse-graining inherently discards geometric information. In atomistic graphs, interatomic distances and bond angles are preserved as edge features; in coarse-grained graphs, these are replaced by building-block connectivity, and the explicit three-dimensional geometry of the framework pore sizes, channel networks, void topology is lost. Whether and how to reintroduce geometric information into coarse-grained MOF models has remained an open question. Our work addresses this through a systematic three-level ablation: pairwise metal–linker distances (RBF edge features), line-graph bond angles, and global persistent-homology descriptors of the pore space.

2.3 Geometric and Topological Descriptors for MOFs

Traditional approaches to MOF property prediction have relied on scalar geometric descriptors computed by software such as Zeo++ [11]: pore-limiting diameter (PLD), largest cavity diameter (LCD), void fraction, gravimetric surface area, and accessible pore volume. These descriptors compress the full three-dimensional pore structure into a handful of numbers and have been widely used as inputs to classical machine learning models for gas storage and separation screening [1, 12, 13]. While effective for ranking and screening, scalar descriptors are limited in their ability to distinguish frameworks with similar pore sizes but different channel topologies or cavity shapes.

Persistent homology (PH) [8] offers a richer geometric representation by characterizing the topological structure of the pore space across all length scales. In the context of MOFs, PH captures the distribution of 1-dimensional features (channels) and 2-dimensional features (voids) as persistence diagrams, which can be vectorized into persistence images for use as model inputs. Krishnapriyan et al. [7] demonstrated that persistence-image features combined with matscholar element embeddings (T+WE), fed to a random forest, substantially outperformed scalar structural descriptors (PLD, void fraction, surface area) for CO_2, CH_4, and N_2 adsorption prediction on the hMOF database. They further showed that different topological features dominate at different pressures: 1D channels are most predictive at low pressure (where pore accessibility governs uptake), while 2D voids dominate at high pressure (where total pore volume governs uptake). However, their analysis was limited to classical machine learning models (random forest, kernel ridge regression) and did not address whether these findings transfer to graph neural network architectures, where the model can potentially learn geometric relationships through message passing.

Our work bridges this gap. We show that persistent-homology features outperform pairwise and angular geometry in a GNN framework, reducing hMOF adsorption error by 53–73% relative to the coarse-grained baseline exceeding both the pairwise-geometry variants (−20% to −47%)

and the Krishnapriyan et al. random forest (−29% to −60%). We further demonstrate that the PH advantage is dataset-dependent: it vanishes on QMOF, where building-block identity fixes geometry, confirming that the topology signal is governed by dataset geometric diversity rather than model capacity.

2.4 MOF Databases and Computational Screening

The scale of the MOF chemical space has motivated the construction of large computational databases. The hMOF database [1] was generated by enumerating 137,953 hypothetical MOFs from 102 building blocks across 6 topologies, with Grand Canonical Monte Carlo (GCMC)-simulated adsorption properties (CO_2, CH_4, N_2) at multiple pressures. The CoRE MOF database [12] curates experimentally synthesized structures from the Cambridge Structural Database, with ~14,000 free-solvent-removed frameworks. The QMOF database [2] provides DFT-computed properties (band gaps, energies, densities of states) for over 20,000 synthesized MOFs, enabling data-driven screening for electronic and catalytic applications.

These databases differ fundamentally in their geometric characteristics. hMOF, as a computational enumeration, varies linker lengths and metal–linker distances independently of chemistry, producing a structurally diverse set with large pores (median PLD 6.75 Å, 80% porous). QMOF, comprising DFT-relaxed synthesized structures, is dominated by dense, small-pore frameworks (median PLD 1.33 Å, 46% with PLD ≤ 1.25 Å) where building-block identity essentially fixes topology. This distinction which we quantify in Section 4.4 is the key to understanding why geometric and topological features help on hMOF but are neutral on QMOF: the utility of geometric representations depends on whether continuum geometry varies independently of building-block identity, which is a property of the dataset, not the target.

3 Methodology

3.1 Datasets and Target Properties

We evaluate all models on two MOF databases that span fundamentally different geometric regimes (Table 1).

QMOF [2] comprises 20,146 DFT-relaxed, experimentally synthesized MOFs with electronic and thermodynamic properties computed at the PBE level. We evaluate three targets: PBE band gap (E_g, eV), energy above hull (E_{hull}, eV/atom), and pore-limiting diameter (PLD, Å). QMOF structures are predominantly dense and small-pore the median PLD is 1.33 Å, 46% of structures have PLD ≤ 1.25 Å, and building-block identity (mofkey) fixes framework topology in all 7,154 mofkey groups. We use this dataset as a building-block-dominated control where geometric features are expected to be redundant with node embeddings.

hMOF GA-51k is a reduced subset of the hypothetical MOF database [1], comprising 50,640 computationally generated structures (of 51,163 total in the GA enumeration, after excluding 523 structures that failed geometry parsing) from a genetic-algorithm enumeration that varies linker length and metal–linker distances independently of chemistry. The dataset is predominantly porous 79.6% of structures have void fraction > 0.5, 81.1% have surface area > 1000 m^2/g, and the median PLD is 6.75 Å with 41% of structures exceeding the 8 Å atomistic cutoff used by CGCNN and ALIGNN. We evaluate six targets: CO_2 uptake at 0.1, 0.5, and 2.5 bar (mol/kg), CH_4 uptake at 2.5 bar (mol/kg), isosteric heat of CO_2 adsorption at 0.5 bar (kJ/mol), and PLD (Å). All adsorption properties were computed via Grand Canonical Monte Carlo (GCMC) simulation at 298 K.

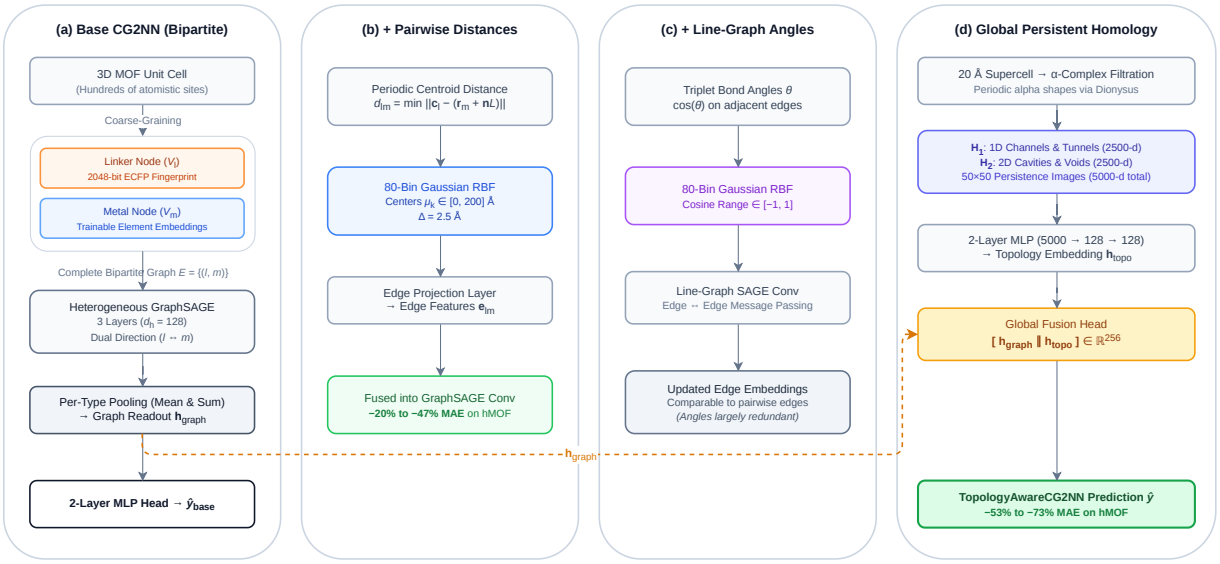


Figure 1 **TopologyAwareCG2NN architecture and representation hierarchy. (a)** Base coarse-grained model (CG2NN) representing MOFs as complete bipartite graphs between linker nodes ($\mathcal{V}_l$, 2048-bit ECFP fingerprints) and metal nodes ($\mathcal{V}_m$, trainable embeddings) processed via heterogeneous GraphSAGE. **(b)** Augmentation with periodic centroid pairwise distances (d_{lm}) expanded over an 80-bin Gaussian RBF ($\Delta = 2.5$ Å). **(c)** Higher-order triplet bond angles (θ) on the edge line graph. **(d)** Global topological feature extraction via periodic α-complex filtration yielding 5000-d persistence images ($\mathcal{H}_1$ channels, $\mathcal{H}_2$ cavities), fused with the graph readout $\mathbf{h}_{\text{graph}}$ into the multimodal latent space $[\mathbf{h}_{\text{graph}} \,\|\, \mathbf{h}_{\text{topo}}] \in \mathbb{R}^{256}$.

Table 1 Target distributions and transformations.

Dataset	Target	Unit	Transform	N	Median	Range	Skew
QMOF	PBE band gap	eV	none	20,146	1.84	0.0 – 7.5	0.42
QMOF	Energy above hull	eV/atom	none	20,146	0.024	0.0 – 1.15	3.84
QMOF	PLD	Å	$\log(1+y)$	20,146	1.33	0.0 – 44.4	3.14
hMOF	CO_2 @ 0.1 bar	mol/kg	$\log(1+y)$	50,640	0.319	0.0 – 6.23	2.27
hMOF	CO_2 @ 0.5 bar	mol/kg	$\log(1+y)$	50,640	1.384	0.0 – 8.93	1.19
hMOF	CO_2 @ 2.5 bar	mol/kg	$\log(1+y)$	50,640	4.580	0.0 – 16.30	0.51
hMOF	CH_4 @ 2.5 bar	mol/kg	$\log(1+y)$	50,640	1.524	0.0 – 9.00	1.08
hMOF	CO_2 heat @ 0.5 bar	kJ/mol	$\log(1+y)$	50,640	21.67	0.0 – 50.60	0.41
hMOF	PLD	Å	$\log(1+y)$	50,640	6.75	1.25 – 24.75	0.96

3.2 Coarse-Grained Graph Representation (CG2NN)

The base coarse-grained model (CG2NN) represents each MOF as a heterogeneous bipartite graph of building blocks rather than individual atoms (Figure 1a).

Nodes. Each unique organic linker in the unit cell is a linker node, initialized with Morgan/ECFP fingerprints [14] (2048-bit, radius 2) projected through a linear layer to the hidden dimension. Each unique metal element in the secondary building units is a metal node, initialized via trainable element embeddings.

Edges. The graph has complete bipartite connectivity: every linker node connects to every metal node (and vice versa), forming a fully connected bipartite graph $\mathcal{E} = \{(l, m) \mid l \in \mathcal{V}_l, m \in \mathcal{V}_m\} \cup \{(m, l)\}$. This captures which building blocks are present but not their spatial arrangement.

Message passing. We use $K = 3$ heterogeneous GraphSAGE convolution layers [15] (hidden dimension $d_h = 128$) with dual-direction message propagation ($l \to m$ and $m \to l$), per-node-type LayerNorm, and ELU activations.

Readout. Per-node-type aggregation (mean and sum pooling over linker and metal node sets separately) is concatenated and passed through a 2-layer MLP head. The base model has 474,565 parameters.

3.3 Geometric Augmentations

We augment the base CG2NN graph with two levels of explicit geometric information (Figure 1b–c).

3.3.1 Pairwise Edge Distances

For each linker–metal edge (l, m), we compute the periodic minimum-image distance between the linker centroid c_l (spatial centroid of all atoms in the molecular fragment) and the metal mean position r_m:

$$d_{lm} = \min_{\mathbf{n}\in\mathbb{Z}^3} \|c_l - (r_m + \mathbf{nL})\|_2 \tag{1}$$

where $\mathbf{L}$ is the lattice matrix. The distance is encoded as an 80-bin Gaussian radial basis function:

$$\mathbf{e}_{lm}^{(k)} = \exp\left(-\frac{(d_{lm} - \mu_k)^2}{2\sigma^2}\right), \quad k \in \{1, \ldots, 80\} \tag{2}$$

with centers μ_k linearly spanning $[0, 200]$ Å at 2.5 Å spacing ($\Delta = 200/80 = 2.5$ Å), under the periodic minimum-image convention. Edge features are projected and fused into the SAGE convolution via an edge projection layer. The edges model has 556,537 parameters.

3.3.2 Line-Graph Bond Angles

For each pair of edges (e_{l,m_1}, e_{l,m_2}) sharing a common node, we compute the angle θ via the cosine of the connecting vectors. The cosine value is expanded into an 80-bin Gaussian RBF over $[-1, 1]$. A line-graph SAGE convolution updates edge embeddings from the angle features before they are aggregated into the main graph convolution. The angles model has 747,001 parameters.

3.4 Persistent-Homology Topology Features

To capture global pore-space geometry channel networks, cavity shapes, and void topology we compute persistent homology (PH) descriptors using continuous alpha-complex filtrations [16] (Figure 1d), following the framework of Krishnapriyan et al. [7].

Supercell. Each structure is expanded to a 20 Å cubic supercell to capture channel loops across periodic boundaries while bounding computation time (~0.1–0.5 s per structure).

Filtration. Periodic alpha shapes are computed via `diode.fill_alpha_shapes` and the `dionysus` persistence engine [17]. We extract 1-dimensional homology ($\mathcal{H}_1$, channels/tunnels) and 2-dimensional homology ($\mathcal{H}_2$, enclosed cavities/voids) persistence barcodes.

Persistence images. Each barcode set is vectorized into a 50×50 persistence image using a linear ramp weighting $w(b, p) = p/p_{\max}$ and an isotropic Gaussian kernel ($\sigma = 0.15$), with birth and persistence axes spanning $[0, 5]$ $\text{Å}^{1/2}$. The two images (1D and 2D) are concatenated into a 5000-dimensional descriptor (2500-d per dimension).

Model integration (TopologyAwareCG2NN). The PH descriptor is projected through a 2-layer MLP ($5000 \to 128 \to 128$, ReLU) and concatenated with the graph readout before the final prediction head:

$$\mathbf{h}_{\text{topo}} = \text{ReLU}(\mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1\mathbf{x}_{\text{topo}} + \mathbf{b}_1) + \mathbf{b}_2) \tag{3}$$

$$\hat{y} = \mathrm{MLP}([\mathbf{h}_{\mathrm{graph}} \,\|\, \mathbf{h}_{\mathrm{topo}}]) \tag{4}$$

The TopologyAwareCG2NN has 1.15M parameters. We note that the PH features are computed as graph-level (global) descriptors and do not participate in message passing they are fused only at the readout stage.

3.5 Baseline Models

All baselines are evaluated on identical train/validation splits with identical structure parity (the same `failed_ids.json` exclusion list), ensuring bit-identical data subsets across all models.

CGCNN [3]. Atomistic crystal graph with 92-dimensional one-hot element embeddings, 8.0 Å cutoff radius, maximum 12 nearest neighbors, and 80-bin Gaussian distance expansion. Three convolutional layers ($d_h = 128$) with atom-wise average pooling.

MEGNet [4] via MatGL v4.0.3. Three interaction blocks with Gaussian bond expansion (100 centers, 0–5 Å, $\sigma = 0.5$), set2set readout, and softplus2 activations (188K parameters).

ALIGNN [5]. Dual-graph architecture: atomistic bond graph $\mathcal{G}$ and interatomic bond-angle line graph $\mathcal{L}(\mathcal{G})$. Four ALIGNN convolution layers coupled with four GCN layers ($d_h = 256$), atom-wise average pooling (4.2M parameters). Graph construction uses a 12-nearest-neighbor, 8.0 Å cutoff the default ALIGNN configuration. We note that this cutoff is the standard used in the original ALIGNN publication and in Choudhary et al. [9]'s MOF benchmarking study.

Random Forest (RF T+WE). Reproduction of the best classical model from Krishnapriyan et al. [7]: a 500-tree `RandomForestRegressor` on the concatenation of (i) the 5000-d persistence images (identical to Section 3.4) and (ii) 1000-d matscholar element word embeddings (200-d Word2Vec per element, aggregated as mean/std/min/max/range), yielding a 6000-d feature vector.

3.6 Training Protocol

Data splitting. All datasets are partitioned using target-decile stratification: the target values are binned into 10 equi-frequency deciles (after log1p transformation for skewed targets), and within each decile, 80% is assigned to training and 20% to validation (deterministic shuffle, seed 42). This ensures balanced representation across the full target range, preventing evaluation bias on heavily skewed distributions. The hMOF split yields 40,512 training / 10,128 validation structures; the QMOF split yields 16,116 / 4,030.

Target transformation. For targets with skewness $\gamma > 1.0$ (all hMOF targets and QMOF PLD), labels are transformed as $y' = \log(1 + \max(y, 0))$ before z-score standardization (computed on training data only). All reported MAE values are evaluated in original physical units after inverse transformation. PBE band gap and E_{hull} are trained in raw units.

Optimization. Table 2 summarizes the training hyperparameters for each model. All CG2NN variants use Adam (lr = 5×10^{-4}, weight decay 10^{-5}, gradient clip 0.5) with ReduceLROnPlateau scheduling (patience 5, factor 0.5, minimum 10^{-5}) and early stopping (patience 15 on validation loss). Seed-43 repeats on hMOF CO_2 @ 2.5 bar confirm reproducibility ($\Delta < 1.1\%$).

Table 2 Training hyperparameters.

Parameter	CG2NN variants	CGCNN	MEGNet	ALIGNN
Optimizer	Adam	Adam	Adam	AdamW
Learning rate	5×10^{-4}	5×10^{-4}	2×10^{-4}	10^{-3}
Weight decay	10^{-5}	10^{-5}	0	10^{-5}
LR scheduler	ReduceLROnPlateau (pat=5, γ=0.5)	MultiStepLR	ReduceLROnPlateau (pat=5)	OneCycleLR
Batch size	64	64	64	8 (hMOF) / 16 (QMOF)
Max epochs	100	100	100	50
Early stopping	15 (val loss)	15 (val MAE)	15 (val loss)	None

Computational environment. CG2NN variants and CGCNN were trained on NVIDIA GTX 1080 Ti and V100 GPUs. ALIGNN was trained on V100 GPUs (batch size 8 for hMOF due to large GA unit cells). MEGNet was trained via the MatGL v4.0.3 framework (PyTorch Geometric backend). RF was trained on 96-core CPU nodes. All experiments use PyTorch [18]; the CG2NN implementations use DGL [19] for heterogeneous graph operations.

4 Results

4.1 Overview

We evaluate eight models CG2NN (base, +Edges, +Angles, +topology), CGCNN, MEGNet, ALIGNN, and a random forest (RF) on nine targets across two datasets: QMOF (20,146 experimentally-synthesized, DFT-relaxed MOFs; three targets: PBE band gap, energy above hull, PLD) and a reduced hMOF subset (50,640 computationally-generated hypothetical MOFs; six targets: CO_2 uptake at 0.1, 0.5, and 2.5 bar, CH_4 uptake at 2.5 bar, heat of CO_2 adsorption, and PLD). All models are trained under an identical protocol: stratified 80/20 train/validation split by target decile, log1p target transformation for skewed distributions, plateau-based learning-rate annealing (patience 5) with early stopping (patience 15), and seed 42 (with seed-43 repeats for key hMOF targets). Performance is reported as validation MAE in original physical units. All percentage improvements are relative to the CG2NN base model on the same target and dataset.

Tables 3 and 4 present the complete results for QMOF and hMOF, respectively. Complete parity plots across all eight benchmarked architectures on CO_2 uptake at 2.5 bar are shown in Figure 6. We organize the discussion around three central findings: (i) geometric features are neutral on QMOF but strongly beneficial on hMOF, governed by dataset topological diversity; (ii) persistent-homology topology features are the strongest geometric representation on hMOF, but this advantage reverses on QMOF; and (iii) atomistic GNNs with local distance cutoffs excel on QMOF's small-pore, local properties but catastrophically fail on hMOF's large-pore adsorption targets.

4.2 QMOF: Geometry Is Neutral on Building-Block-Driven Properties

On QMOF, neither pairwise edge distances nor line-graph angles improve over the CG2NN baseline on any target (Table 3). RBF-encoded metal–linker distances produce changes of −0.1% (PBE band gap), 0.0% (E_{hull}), and +0.03% (PLD) all within noise. Line-graph bond angles are equally neutral: +1.3% on band gap, −0.6% on E_{hull}, and −3.2% on PLD. The small PLD improvement from angles (−3.2%) is consistent across protocols but marginal.

Table 3 QMOF results (validation MAE, stratified 80/20 split, 20,146 structures). Percentage changes are relative to CG2NN base. Best model per target in **bold**.

Target	Unit	CG2NN	+Edges	+Angles	CGCNN	ALIGNN	Topo	MEGNet	RF
PBE band gap	eV	0.2715	0.2711 (−0.1%)	0.2751 (+1.3%)	0.2768 (+2.0%)	**0.2358** (−13.1%)	0.2917 (+7.4%)	0.3862 (+42.2%)	0.4041 (+48.8%)
E_{hull}	eV/atom	0.0165	0.0165 (0.0%)	0.0164 (−0.6%)	**0.0088** (−46.7%)	0.0396 (+140.0%)	0.0174 (+5.5%)	0.0200 (+21.2%)	0.0338 (+104.8%)
PLD	Å (log1p)	0.7695	0.7697 (+0.03%)	0.7446 (−3.2%)	0.5702 (−25.9%)	**0.5317** (−30.9%)	0.8360 (+8.6%)	0.7384 (−4.0%)	0.6794 (−11.7%)

Importantly, this neutrality is not due to an absence of geometric variance in the data. Among 358 building-block groups (mofkey) with two or more MOFs, 101 groups exhibit within-group PLD standard deviations exceeding 1 Å, with a maximum range of 9.06 Å. The model does not exploit this variance the coarse-grained node and bond representation already carries the predictive information, and the added continuous distance and angle channels provide no additional signal on QMOF's building-block-driven properties.

The mechanism is structural: in QMOF, building-block identity (mofkey) essentially fixes topology zero of 7,154 mofkey groups map to multiple framework topologies. When chemistry determines geometry, explicit geometric features are redundant with the node embeddings.

Among the baselines, CGCNN achieves the best E_{hull} (0.0088 eV/atom, −46.7% vs. base) by leveraging its atomistic bond representation, while ALIGNN achieves the best band gap (0.2358 eV) and PLD (0.5317 Å) we analyze these results in Section 4.7. MEGNet performs poorly on QMOF (PBE +42.2%, E_{hull} +21.2%), likely because its coarse chemistry encoding is poorly suited to building-block-dominated properties.

4.3 hMOF: Geometry Strongly Improves Adsorption and PLD Prediction

On the reduced hMOF dataset, the picture inverts dramatically (Table 4). RBF edge distances reduce MAE by 20–47% on every target, with the magnitude following a clear physical hierarchy: (Figure 4)

- **CO_2 uptake pressure trend** (monotone in pressure): −20.6% at 0.1 bar → −26.5% at 0.5 bar → −27.4% at 2.5 bar. At higher pressure, uptake is increasingly dominated by pore volume rather than specific adsorption sites, making the continuum metal–linker geometry more predictive.
- **Second gas confirmation**: CH_4 at 2.5 bar shows −24.1%, confirming the effect is not CO_2-specific.
- **Heat of adsorption** (−36.6%) shows the strongest pairwise-geometry benefit physically expected, as the isosteric heat is the property most directly coupled to pore-wall geometry and adsorbate–framework interaction potentials.
- **PLD** (−46.7%) shows the largest improvement of all targets, because the metal–linker distance is a direct geometric descriptor of pore size.

Line-graph angles perform comparably to edge distances (−20.3% to −47.3%), with no systematic advantage or disadvantage. This indicates that pairwise distances carry the dominant geometric signal, and the additional angular information is largely redundant.

Seed robustness is confirmed: on CO_2@2.5 bar, seed-43 repeats yield edges MAE 1.1420 (Δ 0.3% vs. seed 42) and angles MAE 1.1343 (Δ 1.1%), confirming the results are not seed-dependent. An edge-feature resolution ablation (40 vs. 80 RBF bins) on CO_2@2.5 bar shows no difference (1.2156 vs. 1.2133), proving the benefit arises from geometric information rather than added model capacity.

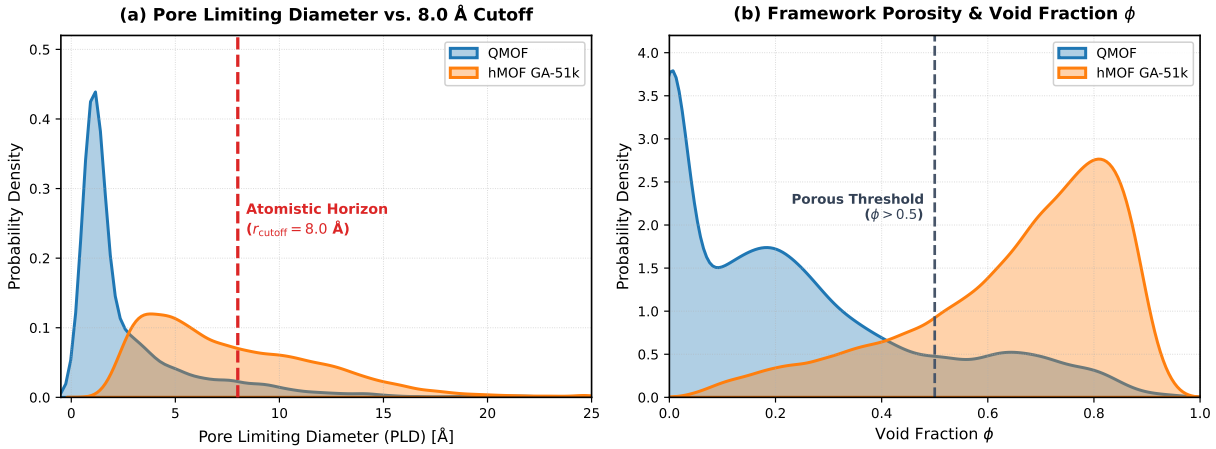


Figure 2 **Pore-limiting diameter (PLD) distributions for QMOF and hMOF.** The two datasets occupy nearly disjoint PLD regimes: QMOF is dominated by dense, small-pore, building-block-driven frameworks (median PLD 1.33 Å), while hMOF spans a broad, continuously varying large-pore regime (median PLD 6.75 Å) generated independently of building-block chemistry.

Table 4 hMOF GA-51k results (validation MAE, stratified 80/20 split, 50,640 structures). Percentage changes are relative to CG2NN base. Best model per target in **bold**.

Target	Unit	CG2NN	+Edges	+Angles	CGCNN	ALIGNN	Topo	MEGNet	RF
CO_2 @ 0.1 bar	mol/kg	0.3376	0.2682 (−20.6%)	0.2690 (−20.3%)	0.1958 (−42.0%)	0.9061 (+168.4%)	**0.1594** (−52.8%)	0.2368 (−29.9%)	0.2383 (−29.4%)
CO_2 @ 0.5 bar	mol/kg	0.7911	0.5816 (−26.5%)	0.5781 (−26.9%)	0.4712 (−40.4%)	1.9590 (+147.6%)	**0.3268** (−58.7%)	0.5563 (−29.7%)	0.5072 (−35.9%)
CO_2 @ 2.5 bar	mol/kg	1.5694	1.1388 (−27.4%)	1.1471 (−26.9%)	1.0194 (−35.0%)	4.3068 (+174.4%)	**0.5942** (−62.1%)	1.2570 (−19.9%)	0.9609 (−38.8%)
CH_4 @ 2.5 bar	mol/kg	0.4990	0.3787 (−24.1%)	0.3787 (−24.1%)	0.3420 (−31.5%)	2.0680 (+314.4%)	**0.2144** (−57.0%)	0.4064 (−18.6%)	0.3433 (−31.2%)
CO_2 heat @ 0.5 bar	kJ/mol	4.1279	2.6151 (−36.6%)	2.6186 (−36.6%)	1.8093 (−56.2%)	12.1221 (+193.7%)	**1.2675** (−69.3%)	2.3783 (−42.4%)	2.0610 (−50.1%)
PLD	Å	2.9089	1.5516 (−46.7%)	1.5334 (−47.3%)	1.6580 (−43.0%)	5.8058 (+99.6%)	**0.7831** (−73.1%)	2.0837 (−28.4%)	1.1717 (−59.7%)

The contrast between QMOF (neutral) and hMOF (strongly positive) on the same geometric features is the central finding of this work. We examine its sharpest manifestation PLD in the next section.

4.4 The Cross-Dataset PLD Confound

PLD is the same physical quantity (pore-limiting diameter, in Å) on both datasets, yet geometric features produce opposite effects:

Table 5 Cross-dataset PLD comparison. The same target shows opposite geometry-feature effects on QMOF vs. hMOF.

	QMOF	hMOF GA
Base MAE (Å)	0.7695	2.9089
+Edges Δ vs. base	+0.03% (neutral)	**−46.7%**
+Angles Δ vs. base	−3.2% (marginal)	**−47.3%**

This is not a target-property effect it is a dataset-property effect. The two datasets occupy nearly disjoint PLD regimes (Figure 2, Table 6):

Table 6 PLD regime statistics for QMOF and hMOF.

Property	QMOF	hMOF GA
Median PLD	1.33 Å	6.75 Å
Mean PLD	2.96 Å	7.76 Å
Range	0 – 44.4 Å	1.25 – 24.75 Å
Skew	3.15	0.96

QMOF comprises DFT-relaxed, experimentally-synthesized MOFs mostly dense, small-pore frameworks where 46% have PLD ≤ 1.25 Å (the hMOF minimum). Building-block identity fixes topology (0/7,154 mofkey groups map to multiple topologies) and largely determines PLD (median within-group PLD std = 0.38 Å). The geometric features are redundant with the node embeddings.

hMOF, by contrast, is a computational genetic-algorithm enumeration that varies linker length and metal–linker distances independently of chemistry. PLD spans 1.25–24.75 Å with 79.6% of structures having void fraction > 0.5 and 81.1% having surface area > 1000 m^2/g. The metal–linker RBF distance is a direct descriptor of pore size the same quantity PLD measures so the geometric features carry genuinely new, predictive information.

This establishes the paper's central mechanistic claim: **geometric features help exactly when continuum geometry varies independently of building-block identity, which is a property of the dataset, not the target.**

4.5 Persistent Homology: The Strongest Geometry Signal on hMOF

Replacing pairwise RBF distances and line-graph angles with global persistent-homology (PH) descriptors of the pore space yields the strongest model on every hMOF target (Table 4, "Topo" column). The TopologyAwareCG2NN model which concatenates 5000-dimensional persistence images (2500-d for 1D channels + 2500-d for 2D voids) with the CG2NN graph readout achieves:

Table 7 Persistent-homology results on hMOF. The Topo model outperforms all other variants on every target.

Target	CG2NN	Topo	Improvement
CO_2 @ 0.1 bar	0.3376	**0.1594**	−52.8%
CO_2 @ 0.5 bar	0.7911	**0.3268**	−58.7%
CO_2 @ 2.5 bar	1.5694	**0.5942**	−62.1%
CH_4 @ 2.5 bar	0.4990	**0.2144**	−57.0%
CO_2 heat @ 0.5 bar	4.1279	**1.2675**	−69.3%
PLD	2.9089	**0.7831**	−73.1%

The PH features outperform not only the pairwise geometry variants but also the atomistic CGCNN model (which itself beats base/edges/angles on every adsorption target). The improvement hierarchy is physically interpretable: the largest gains are on PLD (−73.1%, a direct geometric quantity), heat of adsorption (−69.3%, the most pore-geometry-coupled thermodynamic property), and high-pressure CO_2 uptake (−62.1% at 2.5 bar, where pore volume dominates). The smallest gain is on low-pressure CO_2 (−52.8% at 0.1 bar, where specific adsorption-site chemistry contributes more relative to pore geometry).

The PH features also outperform the random forest model of Krishnapriyan et al. [7], which uses the identical persistence-image features combined with matscholar word embeddings (RF: −29.4% to −59.7% vs. base). The deep TopologyAwareCG2NN extracts more predictive signal from the same topological descriptors than the 500-tree ensemble, confirming that the advantage lies in the features themselves global pore-space topology rather than the model class.

4.6 The Topology Reversal: hMOF vs. QMOF

The most striking result is that the PH topology features, which dominate on hMOF, are **neutral to harmful on QMOF**:

Table 8 Topology feature effect across datasets. The same features and model produce opposite effects.

Target	Dataset	Topo Δ vs. CG2NN
PLD	hMOF	**−73.1%**
PLD	QMOF	**+8.6%**
Band gap	QMOF	+7.4%
E_{hull}	QMOF	+5.5%

The same 5000-dimensional persistence images, processed by the same model architecture, produce a −73% improvement on one dataset and a +8.6% degradation on the other for the same physical target (PLD). This confirms that the topology signal is governed by dataset geometric diversity, not model capacity or target type. When building-block identity fixes geometry (QMOF), the persistence images add noise without information; when continuum geometry varies independently of chemistry (hMOF), they carry the decisive signal.

The random forest reproduces this reversal exactly: RF beats base CG2NN by −29% to −60% on every hMOF target but is worse than base on QMOF's electronic targets (+48.8% to +104.8%) and marginally better on QMOF PLD (−11.7%). This demonstrates that the hMOF-vs.-QMOF contrast is feature-driven, not model-class-driven: the same persistence-image + embedding features help where pore geometry varies independently of chemistry and hurt where it does not, regardless of whether the regressor is a 500-tree random forest or a 1.15M-parameter GNN.

4.7 Atomistic GNNs: Local Cutoff vs. Global Pore Geometry

The atomistic baselines reveal a sharp dichotomy governed by the relationship between pore size and the model's receptive field.

ALIGNN on QMOF best in class. ALIGNN achieves state-of-the-art performance on QMOF band gap (0.2358 eV, −13.1% vs. base) and PLD (0.5317 Å, −30.9% vs. base), outperforming all coarse-grained and atomistic models. This success is explained by pore-size statistics: 91.5% of QMOF pores are ≤ 8.0 Å, meaning ALIGNN's 8 Å atomistic bond graph directly spans the pore windows. On these local, building-block-dominated properties, ALIGNN's explicit bond-angle representation captures the full geometric signal.

ALIGNN on hMOF catastrophic failure. On hMOF, ALIGNN underperforms the coarse-grained CG2NN baseline by +99.6% to +314.4% on every target (Table 4). The mechanism is the receptive-field limitation identified by Choudhary et al. (2022): as illustrated in Figure 3b, with 41% of hMOF structures having PLD > 8 Å (median 6.75 Å, maximum 24.75 Å), the 8 Å atomistic graph connects wall atoms to other wall atoms but cannot bridge across the internal void. The resulting local coordination environment is indistinguishable across pore sizes, and atom-wise average pooling collapses this into a representation that carries no pore-volume information.

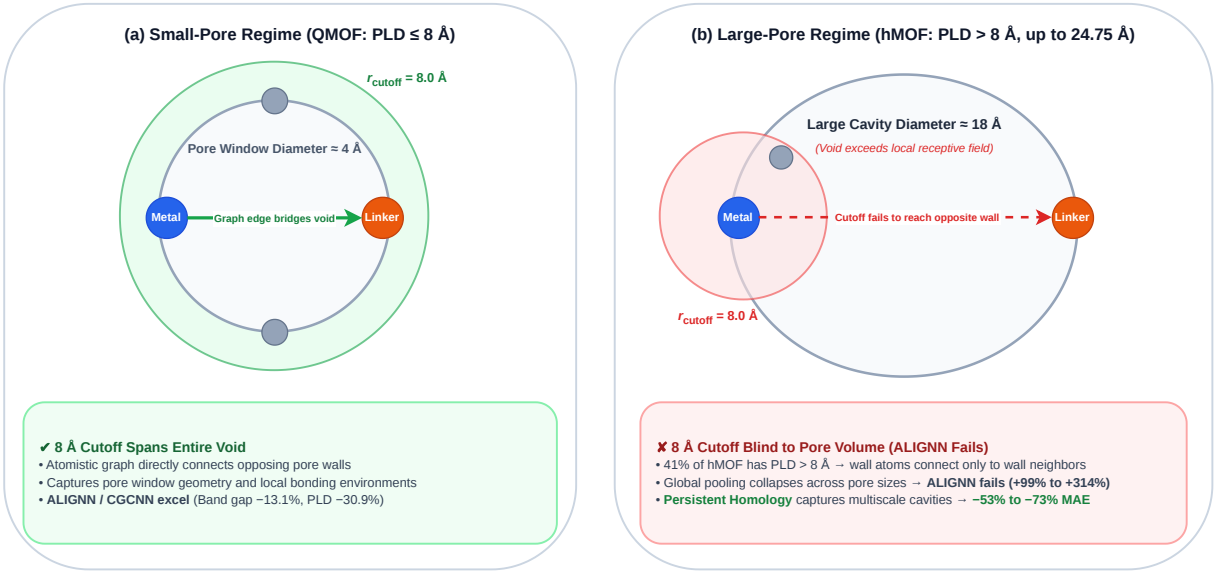


Figure 3 **The cutoff receptive-field dilemma across pore regimes. (a) Small-Pore Regime (**PLD ≤ 8 **Å, e.g., QMOF):** An 8 Å local atomistic cutoff sphere completely spans opposing pore walls, directly capturing pore window geometry and local bonding environments, allowing models like ALIGNN and CGCNN to excel. **(b) Large-Pore Regime (**PLD > 8 **Å, up to** 24.75 **Å, e.g., hMOF):** The 8 Å cutoff sphere connects wall atoms only to adjacent wall neighbors, remaining blind to the interior void volume. Atom-wise global pooling collapses these identical local coordination environments across pore sizes, causing atomistic models to fail while multiscale Persistent Homology captures the global cavity hierarchy.

This finding directly contradicts the conclusion of Choudhary et al. (2022) that "adsorption properties are largely determined by short-range interactions." Their interpretation was based on reasonable ALIGNN performance on the full 137k hMOF database; our results on the reduced GA-51k subset where large-pore structures are more prevalent show that ALIGNN fails catastrophically on adsorption and that global topological features (−53% to −73%) are the decisive signal. The discrepancy likely arises from differences in dataset composition: Choudhary et al. themselves noted that ALIGNN's errors concentrate on high-void-fraction MOFs and that increasing the cutoff to 20 Å yielded no improvement, consistent with our interpretation that the local-graph representation is fundamentally inadequate for large-pore adsorption.

CGCNN follows the same pattern with less severity. It is the best model on QMOF E_{hull} (0.0088 eV/atom, −46.7% vs. base) and competitive on QMOF band gap and PLD. On hMOF, CGCNN beats the coarse-grained base/edges/angles on every adsorption target (−31.5% to −56.2%), though it trails the edge/angle variants on PLD (−43.0% vs. −46.7%/−47.3%) its simpler architecture and Gaussian distance expansion partially compensate for the cutoff limitation but is uniformly worse than the topology-augmented CG2NN. The ranking on hMOF adsorption targets is: Topo > CGCNN > RF ≈ MEGNet > edges ≈ angles > base ≫ ALIGNN; on PLD, CGCNN trails the edge/angle variants.

MEGNet shows a dataset-dependent pattern: it beats the CG2NN base on all hMOF targets (−18.6% to −42.4%) but is the worst model on QMOF (PBE +42.2%, E_{hull} +21.2%), where its coarse chemistry encoding is poorly suited to building-block-dominated electronic properties.

4.8 Subgroup Analysis: Geometry Benefit Scales with Porosity

The hMOF results exhibit a consistent subgroup pattern: geometric features help most on open, highly porous frameworks and least on dense structures. Partitioning the CO_2@2.5 bar validation

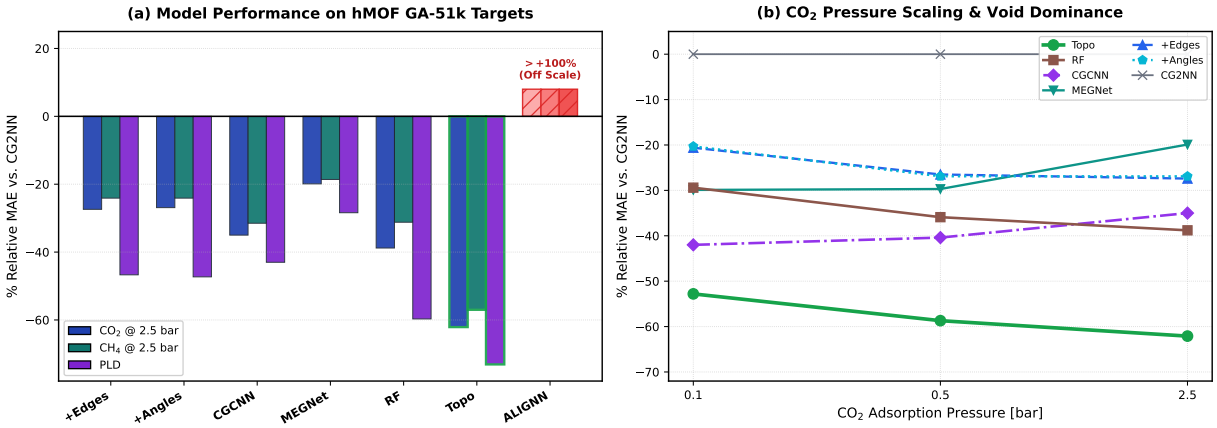


Figure 4 **Benchmark pressure-dependent performance trends on hMOF.** Percentage MAE improvement of the +Edges, +Angles, and Topology-augmented CG2NN variants over the CG2NN base model across the CO_2 adsorption pressure series (0.1, 0.5, 2.5 bar) and CH_4 at 2.5 bar. The improvement magnitude grows monotonically with pressure, consistent with increasing pore-volume dominance of the adsorption mechanism at higher loadings.

set by structural descriptors demonstrates this trend quantitatively (Table 9, Figure 5b): across all three geometric metrics (void fraction ϕ, PLD, and surface area), the relative MAE reduction in the most porous tercile (−41% to −44%) is more than triple that in the densest tercile (−13%).

Table 9 Geometry benefit by porosity subgroup for CO_2 uptake at 2.5 bar (mol/kg).

Subgroup	Edges Δ	Angles Δ
High void fraction (tercile 3)	−41%	−41%
Low void fraction (tercile 1)	−13%	−13%
High PLD (tercile 3)	−43%	−43%
Low PLD (tercile 1)	−26%	−26%
High surface area (tercile 3)	−44%	−44%
Low surface area (tercile 1)	−13%	−13%

By CO_2 uptake decile, the geometry benefit is largest at high-uptake (decile 9: base 3.417 → edges 2.242, −34%) and smallest at low-uptake (decile 0: base 1.561 → edges 1.260, −19%). High-uptake MOFs are precisely those with large, accessible pore volumes the structures where continuum geometry matters most and building-block identity alone is least predictive. This is consistent with the cross-dataset finding: the geometric signal scales with the degree to which the target depends on global pore structure rather than local chemistry. To provide a complete picture across all architectures, Figure 6 presents the validation parity plots for all eight benchmarked models on CO_2 uptake at 2.5 bar, visually illustrating the extreme scatter of ALIGNN and the tight 1:1 correspondence achieved by the topology-augmented CG2NN.

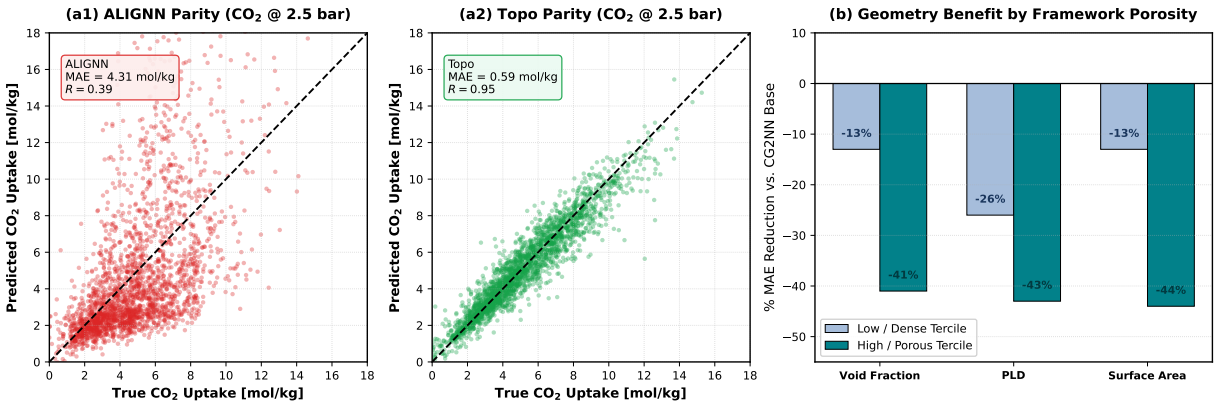


Figure 5 **Predictive parity comparison and porosity subgroup sensitivity on hMOF GA-51k. (a1–a2)** Parity plots of ground-truth vs. predicted CO_2 uptake at 2.5 bar for the worst-performing baseline (ALIGNN, MAE = 4.31 mol/kg, R = 0.39) and the top-performing model (Topo, MAE = 0.59 mol/kg, R = 0.95). Black dashed lines indicate the 1:1 identity. **(b)** Percent MAE reduction achieved by geometric augmentation relative to the base CG2NN model, stratified across low/dense (light blue) and high/porous (teal) terciles for void fraction (ϕ), pore-limiting diameter (PLD), and gravimetric surface area.

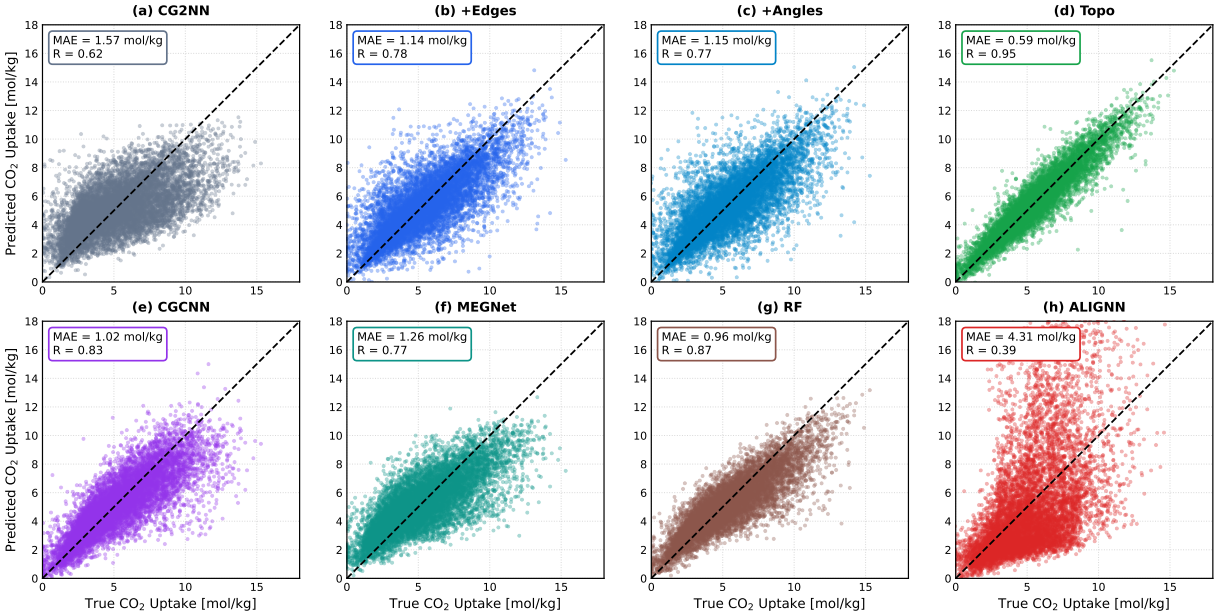


Figure 6 **Validation parity comparison across all eight benchmarked model architectures on hMOF GA-51k (CO_2 @ 2.5 bar). (a–d)** Coarse-grained and geometric variants: (a) CG2NN base, (b) +Edges, (c) +Angles, and (d) Topo (CG2NN + persistent homology). **(e–h)** Atomistic and classical baselines: (e) CGCNN, (f) MEGNet, (g) Random Forest (RF), and (h) ALIGNN. Each panel displays the validation Mean Absolute Error (MAE) and Pearson correlation coefficient (R) computed in linear physical units (mol/kg). Black dashed lines represent the 1:1 identity line.

5 Discussion

5.1 Why Geometry Helps: The Decoupling Principle

The central finding geometric features are neutral on QMOF but decisive on hMOF admits a unifying explanation rooted in the relationship between chemistry and geometry. In any crystalline material, the full structural state is determined by two factors: the chemical identity of building blocks and their spatial arrangement. When these two factors are statistically independent, explicit geometric features carry information not recoverable from node embeddings alone. When they are perfectly correlated when knowing the building blocks deterministically fixes the geometry the node representation is a sufficient statistic, and geometric features are at best redundant and at worst noise.

QMOF exemplifies the correlated regime. This database comprises experimentally synthesized, DFT-relaxed structures where thermodynamic feasibility constrains the structural space: linkers of a given length pair with metals of a given coordination geometry to produce frameworks whose topologies are essentially fixed by chemistry. We verified this directly: zero of 7,154 building-block groups (mofkeys) map to multiple framework topologies, and the median within-group PLD standard deviation is only 0.38 Å. The coarse-grained graph which encodes which building blocks are present and how they connect is already a near-complete description of the structure. Adding distances or topology provides no new information, and the slight degradation from PH features on QMOF (+5.5% to +8.6%) is consistent with the regularization burden of high-dimensional noise features (5000-d) that carry no signal.

hMOF exemplifies the decoupled regime. The genetic-algorithm enumeration systematically varies linker length and metal–linker distance while holding chemistry fixed, producing structures where the same building blocks can yield PLDs ranging from 1.25 to 24.75 Å. Here, the node representation captures *what* is in the structure but not *how it is arranged in space*, and the geometric features supply the missing dimension. The magnitude of the benefit tracks the degree of geometric dependence in the target: PLD (−73.1%) and heat of adsorption (−69.3%), which are directly governed by pore geometry, show the largest gains; low-pressure CO_2 uptake (−52.8%), where adsorption-site chemistry contributes more, shows the smallest.

This decoupling principle is not specific to MOFs. It predicts that geometric representations will be valuable in any materials domain where structural diversity exceeds chemical diversity for example, in zeolite frameworks with fixed composition but variable topology, in polymer packings with fixed monomers but variable conformation, or in alloy surfaces with fixed composition but variable ordering. Conversely, it predicts neutrality in domains like molecular property prediction, where the SMILES string is a complete description of the structure.

5.2 The Length-Scale Dichotomy and Atomistic Model Failure

The failure of ALIGNN on hMOF adsorption errors 99–314% worse than the coarse-grained baseline is not a model capacity problem. ALIGNN has 4.2M parameters, an order of magnitude more than any CG2NN variant, and it achieves state-of-the-art performance on QMOF band gap (0.2358 eV) and PLD (0.5317 Å). The failure is architectural: the 8 Å atomistic bond graph cannot represent the pore geometry that governs adsorption in high-porosity MOFs.

The mechanism is illustrated by a simple geometric argument (Figure 3). In a MOF with PLD = 15 Å, the pore is a void roughly 15 Å in diameter. ALIGNN's 8 Å cutoff sphere centered on any wall atom reaches 8 Å into the void barely past the center and connects only to other wall atoms. The resulting local coordination environment is nearly identical for a 10 Å pore and a 20 Å pore: the same atoms, the same bonds, the same angles. The pore volume the quantity that determines

high-pressure uptake is invisible to the model. Atom-wise average pooling then collapses these indistinguishable local environments into a single vector, washing out any residual signal.

CGCNN, which uses the same 8 Å cutoff but a simpler architecture (no line graph), performs better than ALIGNN on hMOF but still worse than the topology-augmented coarse-grained model. Its Gaussian distance expansion partially compensates by encoding finer-grained distance information, and its lower capacity (fewer parameters) reduces overfitting, but it suffers from the same fundamental receptive-field limitation. The ranking on hMOF adsorption targets – Topo > CGCNN > RF ≈ MEGNet > edges ≈ angles > base ≫ ALIGNN; in PLD, CGCNN trails the edge/angle variants – reflects a hierarchy of receptive fields, not a hierarchy of model capacity.

5.3 Why Persistent Homology Outperforms Pairwise Geometry

The PH topology features (−53% to −73%) substantially outperform pairwise edge distances (−20% to −47%) on every hMOF target, despite both encoding geometric information. The distinction lies in *what aspect* of geometry each representation captures.

Pairwise RBF distances encode the metric distance between building-block centroids a scalar descriptor of linker-to-metal separation. This is informative because, in hMOF, this distance correlates with pore size: longer linkers produce larger pores. But it is a *local* and *pairwise* descriptor: it captures the distance between two specific building blocks without information about how those building blocks are arranged in the full three-dimensional pore network. A MOF with the same set of pairwise metal–linker distances can have fundamentally different pore topologies isolated cages versus connected channels, 1D tunnels versus 3D networks and the pairwise representation cannot distinguish them.

Persistent homology captures exactly this missing information. The 1D homology ($\mathcal{H}_1$) features encode channel-like topological features connected tunnels that thread through the framework while the 2D homology ($\mathcal{H}_2$) features encode enclosed voids and cavities. These are *global* descriptors: they depend on the full pore-space architecture, not on any pairwise distance. The persistence diagram further encodes the *scale* at which each feature exists (birth and death in the alpha-complex filtration), providing a multiscale characterization that naturally captures both small pore windows and large cavity volumes.

This explains the pressure-dependent hierarchy we observe. At low pressure (0.1 bar), CO_2 adsorption is dominated by specific binding sites the strongest adsorption sites fill first and pore accessibility (1D channels) matters more than total volume. The pairwise distance, which correlates with pore size, provides moderate benefit (−20.6%), and PH, which captures channel connectivity, provides larger benefit (−52.8%). At high pressure (2.5 bar), uptake is dominated by total pore volume all accessible void space fills and the 2D homology features, which encode cavity volumes, become most predictive. The PH benefit grows (−62.1%), while the pairwise benefit grows more slowly (−27.4%). This pressure-dependent shift in feature importance was first noted by Krishnapriyan et al. [7] in their random-forest feature-importance analysis; our results confirm it in a GNN framework and show that the effect is preserved when the model can learn feature interactions end-to-end.

5.4 The Same Features, Different Models: Feature-Driven, Not Model-Driven

A striking result is that the hMOF-vs.-QMOF contrast reproduces across fundamentally different model architectures. The RF model a 500-tree random forest on hand-crafted persistence-image and word-embedding features beats the CG2NN base by −29% to −60% on every hMOF target but is worse than base on QMOF's electronic targets (+49% to +105%) and marginally better on QMOF PLD (−12%). The TopologyAwareCG2NN a 1.15M-parameter GNN that learns feature

interactions through message passing shows the same qualitative pattern (−53% to −73% on hMOF, +5% to +9% on QMOF). Two radically different model classes, the same features, the same dataset-dependent contrast.

This demonstrates that the governing factor is the *representation* (what information the features encode), not the *architecture* (how that information is processed). The deep model extracts more signal from the same persistence images outperforming RF by an additional 15–25 percentage points on hMOF but both models agree on *where* the features help and *where* they hurt. This has a practical implication: if one must choose between investing in better features or better architectures for a given application, the dataset's geometric diversity should guide that decision. For structurally diverse databases like hMOF, richer geometric features yield large returns regardless of model class. For chemistry-constrained databases like QMOF, architectural improvements (e.g., ALIGNN's line-graph angles) yield returns that feature engineering cannot match.

5.5 Generalization and Boundaries

The decoupling principle provides a framework for predicting when geometric representations will help in new settings, but several boundaries of our study should be noted.

First, the QMOF/hMOF dichotomy is extreme nearly perfectly correlated versus nearly perfectly decoupled chemistry–geometry relationships. Real-world databases often occupy an intermediate regime. CoRE MOF [12] (experimentally synthesized, ~14k structures) has higher topological diversity than QMOF but lower than hMOF; we would predict moderate geometric benefits there, but the exact magnitude remains to be tested.

Second, our PH features are global descriptors fused only at the readout stage. An alternative design embedding topological information into edge or node features for participation in message passing could potentially capture local-topological interactions that the global descriptor misses. The current design treats topology as a graph-level summary; a more fine-grained integration might yield further gains, particularly for properties that depend on *local* pore geometry (e.g., gas separation selectivity, which depends on pore-window size rather than total volume).

Third, our analysis focuses on regression MAE. The relative rankings of models might differ under other metrics (RMSE, which penalizes large errors more heavily, or classification metrics for screening applications). The ALIGNN failure on hMOF, for instance, is driven partly by catastrophic errors on high-uptake structures; under a classification metric that only cares about ranking, the degradation might be less severe.

6 Limitations & Future Work

6.1 Limitations

Dataset scope. This study evaluates two datasets that represent extreme points on the chemistry–geometry coupling spectrum: QMOF (nearly perfectly correlated) and hMOF GA-51k (nearly perfectly decoupled). While this dichotomy was instrumental for isolating the decoupling principle, most real-world MOF databases occupy an intermediate regime. CoRE MOF [12] (~14k experimentally synthesized structures) has higher topological diversity than QMOF but lower than hMOF; CURATED MOF and ToBaCCo-generated structures similarly span a middle ground. The magnitude of geometric feature benefit in these intermediate regimes and whether the decoupling principle's predictions hold quantitatively remains to be tested. We note that the principle itself is general: it predicts the *direction* of the effect (geometry helps when decoupled, is neutral when correlated) but our results do not constrain the *magnitude* in intermediate regimes.

Single split. All experiments use a single stratified 80/20 split (seed 42). While seed-43 repeats on hMOF CO_2 @ 2.5 bar confirm reproducibility ($\Delta < 1.1\%$), a full k-fold cross-validation would provide tighter confidence intervals on relative model rankings, particularly for the smaller QMOF dataset where the validation set contains 4,030 structures. The close margins on QMOF (e.g., CG2NN base vs. edges on PBE band gap: 0.2715 vs. 0.2711, a 0.15% difference) are within the expected variance of a single split.

Global PH integration. The persistent-homology features are computed as graph-level (global) descriptors and fused only at the readout stage via concatenation with the graph readout. They do not participate in message passing. This design choice was motivated by simplicity and by the fact that PH is inherently a global property (the alpha-complex filtration depends on the full structure). However, an alternative design embedding topological information into edge or node features for participation in message passing could potentially capture local-topological interactions that the global descriptor misses. Properties that depend on *local* pore geometry (e.g., gas separation selectivity, which depends on pore-window size rather than total volume) might benefit from such fine-grained integration.

ALIGNN cutoff not systematically varied. Our analysis of ALIGNN's failure on hMOF attributes it to the 8 Å atomistic graph cutoff being smaller than the pore diameter in 41% of structures. While this interpretation is supported by the geometric argument, we did not systematically vary the cutoff distance. Choudhary et al. [9] reported that increasing the ALIGNN cutoff to 20 Å yielded no improvement, which is consistent with our interpretation (a larger cutoff adds noise from distant, irrelevant atoms without resolving the fundamental receptive-field limitation) but does not definitively rule out a cutoff-dependent effect.

Evaluation metric. All results are reported as MAE in original physical units. The relative rankings might differ under other metrics. RMSE, which penalizes large errors more heavily, would amplify the ALIGNN failure on hMOF (driven by catastrophic errors on high-uptake structures). Classification metrics for screening applications e.g., top-k precision for identifying the highest-uptake structures might show different model rankings, particularly if the models agree on rankings but disagree on absolute values.

PH computation cost. While individual PH computations are fast (~0.1–0.5 s per structure), the 5000-dimensional persistence images add a preprocessing step that atomistic models do not require. For very large databases ($> 10^5$ structures) or for streaming/online evaluation settings, this cost is non-trivial, though it is orders of magnitude cheaper than GCMC simulation.

6.2 Future Work

6.2.1 Representation Transfer to Generative Models for Inverse Design

The most compelling future direction flows directly from our central finding. The TopologyAwareCG2NN learns a representation $\mathbf{h}_{\text{final}} = [\mathbf{h}_{\text{graph}} \| \mathbf{h}_{\text{topo}}] \in \mathbb{R}^{256}$ that fuses building-block identity with global pore topology in a single property-aligned latent space. This representation is:

1. **Coarse-grained** operating at the building-block level, compatible with existing MOF generative models (MOFDiff [20], LVGP-MOBBO [21], SELFIES-based approaches) that generate structures at the linker/metal level.
2. **Differentiable** gradients flow from the property prediction through the encoder to the input features, enabling gradient-based optimization of building-block selections.
3. **Property-aligned** the latent space is trained on adsorption prediction, so adsorption-relevant geometric features (channel connectivity, cavity volume) are encoded as principal axes.

Current MOF generative pipelines suffer from a representation mismatch: the generative model's latent space and the property predictor's representation space are separate, and the predictor must be either too simple (an MLP on the latent code) or misaligned (ALIGNN with its 8 Å cutoff blindness to large pores). A generative model that initializes from or is conditioned on the pretrained TopologyAwareCG2NN encoder would operate in a latent space where adsorption-relevant pore topology is already encoded, addressing this mismatch at its root.

Several architectures could leverage this:

- **VAE with pretrained encoder.** The encoder maps MOF structures to the 256-d latent code $\mathbf{h}_{\text{final}}$; a decoder is trained to reconstruct building-block graphs from this code. Property-guided generation optimizes in the latent space where the property-relevant directions are already learned.
- **Diffusion model with property conditioning.** Building-block graph diffusion conditioned on the pretrained representation, where the conditioning vector carries the topology-aware signal.
- **Latent-space optimization.** Direct gradient-based optimization of the latent code to maximize predicted adsorption, with a decoder mapping optimized codes back to building-block graphs.

The primary challenge is that the PH branch requires 3D crystal structure at inference time, which a building-block-level generative model may not produce. Three bridging strategies are viable: **(i)** using the graph branch representation alone (fully self-contained at the building-block level, with moderate accuracy loss the edges variant still achieves −27% vs. base on hMOF adsorption); **(ii)** training a graph-to-PH predictor to make the full representation self-contained; or **(iii)** integrating 3D structure assembly into the generative loop (as in MOFDiff) to compute PH on generated candidates. The first strategy is immediately implementable; the latter two are concrete directions for future architecture design.

6.2.2 Intermediate-Regime Databases

Extending evaluation to CoRE MOF, CURATED MOF, and ToBaCCo-generated structures would test the quantitative predictions of the decoupling principle in intermediate chemistry–geometry coupling regimes. Of particular interest is whether the benefit of PH features scales continuously with the degree of decoupling (as measured, e.g., by the within-mofkey PLD variance) or whether there is a sharp transition. A continuous scaling would strengthen the principle's predictive utility; a sharp transition would suggest an unmodeled threshold effect.

6.2.3 Local Topological Features via Message Passing

Integrating persistent-homology information into edge or node features rather than as a global readout-level descriptor could capture local pore geometry relevant to properties like separation selectivity. One approach: compute PH on local regions (e.g., per-pore-window) and assign the resulting descriptors to the nearest building-block edges, allowing the message-passing layers to propagate local topological information through the graph. This would require developing a principled local-PH computation for periodic structures, which is an open problem in computational topology.

6.2.4 Differentiable Persistent Homology

Making the alpha-complex filtration differentiable so that gradients flow through the persistence-image computation into the generative model would enable end-to-end training of a topology-aware generative pipeline. Differentiable TDA is an active research area [22, 23], but existing

implementations handle point clouds and image data, not 3D periodic crystal structures. Extending differentiable PH to periodic alpha complexes would remove the need for the bridging strategies discussed in Section 6.2.1 and enable fully end-to-end optimization of pore topology.

6.2.5 Multi-Task and Transfer Learning

All models in this study are trained on single targets. Multi-task training jointly predicting adsorption at multiple pressures, or jointly predicting adsorption and PLD could exploit correlations between targets to improve data efficiency, particularly for the smaller QMOF dataset. Additionally, transfer learning from hMOF (where topology features are informative and the dataset is large) to QMOF (where data is scarcer) could test whether the topology-aware representation transfers across geometric regimes, though the QMOF results suggest the transferred topology features may be noise in the correlated regime.

7 Conclusion

This study systematically evaluated three levels of geometric representation pairwise distances, bond angles, and global persistent-homology topology across two structurally distinct MOF databases and nine property targets, benchmarked against atomistic graph neural networks (CGCNN, MEGNet, ALIGNN) and a classical random-forest baseline.

The findings are governed by a single principle: **geometric features are informative exactly when continuum geometry varies independently of building-block identity.** On QMOF, where building-block identity fixes framework topology (0/7,154 mofkey groups map to multiple topologies), geometric and topological features are uniformly neutral or slightly harmful, and atomistic models with local receptive fields achieve the best performance. On hMOF GA-51k, where a genetic-algorithm enumeration decouples geometry from chemistry, the same features reduce error by 20–73%, with persistent-homology topology features providing the largest gains (−53% to −73%) and outperforming all atomistic models. The same pattern reproduces across a 500-tree random forest and a 1.15M-parameter GNN, confirming that the governing factor is the representation, not the architecture.

These results resolve a contradiction in the literature: adsorption in high-porosity MOFs is governed by global pore topology, not short-range atomic interactions, and atomistic models with fixed local cutoffs are structurally incapable of capturing this signal when pores exceed the cutoff radius. The practical implication is that representation choice should be guided by the dataset's geometric diversity and the property's governing length scale, not by assumptions of universal superiority for any single model class.

References

[1] Christopher E. Wilmer et al. "Large-scale screening of hypothetical metal–organic frameworks". In: *Nature Chemistry* 4.2 (2012), pp. 83–89. DOI: 10.1038/nchem.1192.

[2] Andrew S. Rosen et al. "Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery". In: *Matter* 4.5 (2021), pp. 1578–1597. DOI: 10.1016/j.matt.2021.02.015.

[3] Tian Xie and Jeffrey C. Grossman. "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties". In: *Physical Review Letters* 120.14 (2018), p. 145301. DOI: 10.1103/PhysRevLett.120.145301.

[4] Chi Chen et al. "Graph networks as a universal machine learning framework for molecules and crystals". In: *Chemistry of Materials* 31.9 (2019), pp. 3564–3572. DOI: 10.1021/acs.chemmater.9b01294.

[5] Kamal Choudhary and Brian DeCost. "Atomistic line graph neural network for improved materials property predictions". In: *npj Computational Materials* 7 (2021), p. 185. DOI: 10.1038/s41524-021-00650-1.

[6] Vadim Korolev and Artem Mitrofanov. "Coarse-grained crystal graph neural networks for reticular materials design". In: *Journal of Chemical Information and Modeling* 64.6 (2024), pp. 1919–1931. DOI: 10.1021/acs.jcim.3c02083.

[7] Aditi S. Krishnapriyan et al. "Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal–organic frameworks". In: *Scientific Reports* 11.1 (2021), p. 8888. DOI: 10.1038/s41598-021-88027-8.

[8] Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. Vol. 69. American Mathematical Society, 2010.

[9] Kamal Choudhary et al. "Graph neural network predictions of metal organic framework CO_2 adsorption properties". In: *Computational Materials Science* 211 (2022), p. 111388. DOI: 10.1016/j.commatsci.2022.111388.

[10] Jingqi Wang, Jingbo Liu, Han Wang, et al. "A comprehensive transformer-based approach for high-accuracy gas adsorption predictions in metal-organic frameworks". In: *Nature Communications* 15 (2024), p. 1904. DOI: 10.1038/s41467-024-46276-x.

[11] Thomas F. Willems et al. "Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials". In: *Microporous and Mesoporous Materials* 149 (2012), pp. 134–141. DOI: 10.1016/j.micromeso.2011.08.020.

[12] Yongchul G. Chung et al. "Computation-Ready, Experimental Metal–Organic Frameworks: A Tool To Enable High-Throughput Screening of Nanoporous Crystals". In: *Chemistry of Materials* 26.21 (2014), pp. 6185–6192. DOI: 10.1021/cm502594j.

[13] Yamil J. Colón and Randall Q. Snurr. "High-throughput computational screening of metal-organic frameworks". In: *Chemical Society Reviews* 43.16 (2014), pp. 5735–5749. DOI: 10.1039/c4cs00070f.

[14] David Rogers and Mathieu Hahn. "Extended-connectivity fingerprints". In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754. DOI: 10.1021/ci100050t.

[15] William L. Hamilton, Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[16] Herbert Edelsbrunner and Ernst P. Mücke. "Three-dimensional alpha shapes". In: *ACM Transactions on Graphics* 13.1 (1994), pp. 43–72. DOI: 10.1145/174462.156635.

[17] Dmitriy Morozov. *Homological illusions of persistence and stability*. PhD thesis, Duke University. 2008.

[18] Adam Paszke et al. "PyTorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[19] Minjie Wang et al. "Deep graph library: A graph-centric, highly-performant package for graph neural networks". In: *arXiv preprint arXiv:1909.01315* (2019).

[20] Xiang Fu et al. "MOFDiff: Coarse-grained diffusion for metal–organic framework design". In: *International Conference on Learning Representations (ICLR)*. 2024.

[21] Yigitcan Comlek et al. “Rapid design of top-performing metal-organic frameworks with qualitative representations of building blocks”. In: *npj Computational Materials* 9.1 (2023), pp. 1–9.

[22] Christoph Hofer et al. “Deep learning with topological signatures”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[23] Mathieu Carrière et al. “PersLay: A neural network layer for persistence diagrams and new graph topological signatures”. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. 2021, pp. 2786–2796.