Interaction Topological Transformer for Multiscale Learning in Porous Materials

Dong Chen[1], Jian Liu[1], Chun-Long Chen [†2], and Guo-Wei Wei [*1,3,4]

[1] *Department of Mathematics, Michigan State University, MI, 48824, USA*
[2] *Physical Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99354, United States*
[3] *Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA*
[4] *Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*

Abstract Porous materials exhibit vast structural diversity and support critical applications in gas storage, separations, and catalysis. However, predictive modeling remains challenging due to the multiscale nature of structure–property relationships, where performance is governed by both local chemical environments and global pore-network topology. These complexities, combined with sparse and unevenly distributed labeled data, hinder generalization across material families. We propose the Interaction Topological Transformer (ITT), a unified data-efficient framework that leverages novel interaction topology to capture materials information across multiple scales and multiple levels, including structural, elemental, atomic, and pairwise-elemental organization. ITT extracts scale-aware features that reflect both compositional and relational structure within complex porous frameworks, and integrates them through a built-in Transformer architecture that supports joint reasoning across scales. Trained using a two-stage strategy, i.e., self-supervised pretraining on 0.6 million unlabeled structures followed by supervised fine-tuning, ITT achieves state-of-the-art, accurate, and transferable predictions for adsorption, transport, and stability properties. This framework provides a principled and scalable path for learning-guided discovery in structurally and chemically diverse porous materials.

Keywords Interaction homology, Multiscale topology, Gas selectivity, Diffusivity, Transformer.

[†]Corresponding author: chunlong.chen@pnnl.gov
[*]Corresponding author: weig@msu.edu

Contents

1 Introduction

Porous materials such as metal–organic frameworks (MOFs), covalent organic frameworks (COFs), zeolites, and porous polymers underpin applications in gas storage and separations [1], carbon capture [2], catalysis [3], and sensing [4, 5]. Their technological promise arises from modular building blocks, tunable pore architectures, and rich surface chemistries, which together create an effectively unbounded design space with many structures already synthesized and many more accessible in silico [6, 7]. For example, polyoxometalate-based MOFs (POM-MOFs) represent a particularly versatile class, combining the tunable electronic and redox properties of POMs [8] with the open, modular networks of MOFs. These hybrid materials have shown great potential in heterogeneous catalysis, where the POM units provide functional activity while the MOF framework ensures accessibility and structural tunability [9], exemplifying how functional building blocks and networked architectures can be integrated to advance porous-material applications [10].

However, accurate prediction remains difficult because performance depends not only on local surface chemistry but also on the long-range topology and connectivity of the pore network, which complicates faithful and efficient simulation across conditions and tasks; conducting independent virtual screenings for each application is impractical at scale [11, 12]. More broadly, property prediction in porous materials remains challenging because structure–property relationships are inherently multiscale: behavior emerges from the interplay between local chemical environments and the global geometry and connectivity of pore networks. Performance is also highly sensitive to operating conditions, such as temperature, pressure, and adsorbate identity, with long-range, network-mediated effects often outweighing purely local descriptors [13, 14]. Compounding these issues, labeled data are limited and uneven across material families and target properties, and are generated under diverse experimental and computational protocols, which complicates supervised learning and hinders generalization [15]. These realities motivate learning approaches that integrate local chemistry with pore-network topology while remaining data-efficient and transferable across porous-material families and operating conditions.

Prior work spans physics-informed descriptors [16, 17], such as surface area, pore-size distributions, and chemical fingerprints, as well as learned representations based on graph and geometric neural networks (GNNs) originally developed for molecules and dense crystals [18, 19]. In parallel, math-based approaches [20, 21, 22] have been explored to incorporate mathematical principles into porous materials modeling. While these methods face limitations in porous-materials settings: hand-crafted descriptors only partially encode void topology and channel connectivity; local message passing under-represents higher-order (multi-body) interactions [22] and long-range effects transmitted through connected pores; and family-specific models transfer poorly across MOFs, COFs, zeolites, and porous polymers [23]. In practice, realistic frameworks often require multiple periodic images or large supercells—frequently exceeding 10^4 atoms—to capture pore networks and guest–framework interactions, rendering conventional GNNs computationally prohibitive and memory-bound. Voxelization and other 3D-grid approaches convert structures into images to leverage computer-vision architectures [24], but they sacrifice key invariances (rotation, translation, permutation), suffer from extreme sparsity and dimensionality, and complicate faithful treatment of periodicity. Moreover, existing pretraining strategies commonly rely on engineered priors [23, 25], such as connectivity types or hand-computed void metrics, as supervision, which limits scalability and can bias the learned representation rather than allowing the model to discover transferable

topological structure directly from raw data.

Here, we introduce interaction topology [26, 27] as a novel multilevel interactive representation that integrates structural, elemental, and pairwise elemental information within a single formalism. Unlike classical persistent homology (PH) [28], which analyzes a system as a whole, interaction topology examines how element-defined subsets contribute both individually and jointly. We instantiate this via an interaction complex (IC), a generalization of the simplicial complex that encodes higher-order and element-specific information. Applying persistent interaction homology (PIH) to this complex yields scale-aware summaries at three complementary levels: form individual clusters, cluster interactions, to global pore architecture. In parallel, we preserve an atomic view: a Crystal Graph Convolutional Neural Network (CGCNN) [18] module (without global pooling) produces locality-preserving atomic embeddings, ensuring that fine-scale chemical and geometric signals are retained rather than collapsed. Building on these ingredients, the Interaction Topological Transformer (ITT) fuses four embedding streams, including structural, elemental, atomic, and pairwise interaction, into a unified model. A transformer attends across all tokens so that topology-derived patterns can reinforce or counterbalance localized chemical environments. Cross-attention layers inject interaction-topology signals into the multilevel pathway, encouraging the network to integrate long-range pore-network dependencies with local information. The CGCNN branch refines atomic tokens before fusion, while the topology channels contribute persistence-based descriptors of connected components, tunnels, and cavities across scales. After multiple ITT layers, property-specific output heads support adsorption, transport, and stability queries within the ITT architecture.

Furthermore, we adopt a two-stage training strategy. During pretraining, ITT is exposed to a large unlabeled corpus (over 0.6 million structures) spanning MOFs, COFs, zeolites, porous polymers, and other pore-network materials—without external priors, auxiliary data, or labels. Self-supervised learning recovers masked structural-level embeddings from the remaining multilevel context (atomic, elemental, and interaction-topology signals), encouraging alignment between geometric, topological, and chemical views. In fine-tuning, the pretrained backbone is adapted to labeled benchmarks covering a variety of properties—including adsorption (Henry coefficients, uptakes, selectivities)[29, 30], transport (diffusivities)[31], and stability/mechanical responses [32]. Across these benchmarks, ITT delivers accurate, transferable, and family-agnostic predictions, outperforming previous baselines and improving cross-family generalization. The multilevel design also provides interpretability: attention concentrated on high-persistence structural and interaction features aligns with known adsorption sites and rate-limiting apertures, while atom-level signals remain traceable through the CGCNN branch. Taken together, the proposed interaction topological Transformer forms a coherent framework that reconciles local chemistry with pore-network topology for data-efficient prediction across diverse porous materials.

2   Results

2.1   Interaction Topological Transformer

A central contribution of the proposed ITT framework is its integration of persistent interaction homology as a topological foundation for multilevel material representation. While traditional graph-based models are limited to pairwise atomic interactions, and classical PH focuses on global topology without differentiating atom types, PIH advances this landscape by introducing interaction
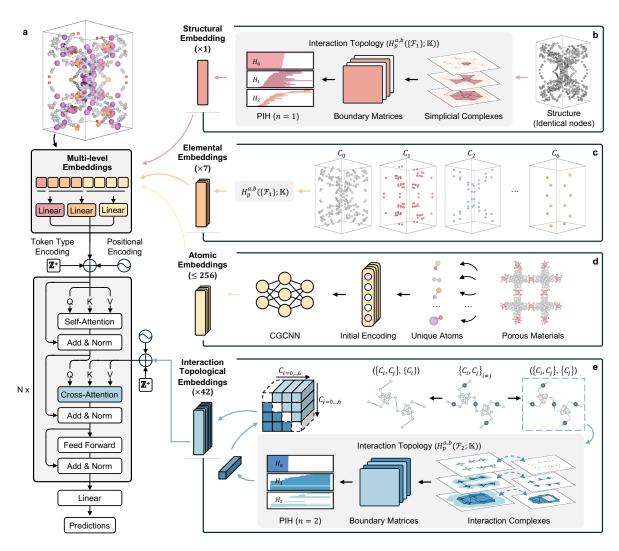
Figure 1: Overall framework of the Interaction Topological Transformer (ITT). **a** Given a porous material structure, ITT constructs multi-level embeddings consisting of structural-, elemental-, atomic-, and interaction-topological–level representations. **b** Structural embedding: the entire structure is modeled as simplicial complexes where all atoms are treated as identical nodes. Persistent interaction homology (PIH, $n = 1$) is applied to extract topological features, yielding a single structural token. **c** Elemental embeddings: atoms are grouped into seven element clusters (C_0–C_6), and PIH ($n = 1$) is computed for each cluster, producing seven tokens that represent cluster-specific topological contributions. **d** Atomic embeddings: a CGCNN without pooling encodes unique atoms, generating locality-preserving atomic representations with up to 256 atomic tokens. **e** Interaction topological embeddings: element-cluster interactions are represented by interaction simplicial complexes, where PIH ($n = 2$) captures higher-order relations across clusters. Considering 7 element clusters, 42 pairwise interaction tokens are derived. Together, these multi-level embeddings are fused through self-attention and cross-attention layers to form the final predictive representation of porous material properties.

topology-a formalism designed to capture structural, elemental, and cross-elemental organization within a unified framework [26, 27]. At its core, PIH operates on interaction complexes (ICs), which are a generalizations of simplicial complexes that encode both intra- and inter-cluster geometric relationships. By applying PIH to these complexes, the ITT model generates scale-aware descriptors at three complementary levels: (i) a structural-level embedding summarizing the global

pore geometry; (ii) elemental-level embeddings based on clusters derived from co-occurrence and chemical similarity; and (iii) pairwise interaction-level embeddings that characterize the topological relationships between elemental clusters (e.g., S–Tb interactions). Figure 4**c**–**f** illustrate how PIH captures both conventional features (e.g., connected components, tunnels, voids) and emergent, chemically informed interaction motifs that classical PH would overlook.

Importantly, PIH enables the model to retain local chemical detail while also learning from inter-element interactions, a capability demonstrated in the comparison of different interaction centers in Figure 4**e**–**f**. For example, choosing S versus Tb as the interaction center yields qualitatively different H_0 and H_1 topological patterns—capturing, respectively, the number of centers and the multi-scale formation of triangle-like structures. Through multiscale filtrations (e.g., $\alpha = 1$, 2, 2.5 Å), PIH systematically quantifies the evolution of interaction geometry, preserving both fine-grained local features and global connectivity. This interaction-centered perspective is critical for modeling porous materials, where function often emerges from the joint behavior of chemically distinct subsystems.

Building upon this topological foundation, ITT fuses four embedding streams—structural, elemental, atomic, and pairwise interaction—into a unified Transformer architecture. The topology-derived embeddings are projected and fed as discrete tokens into the Transformer, where cross-attention layers inject interaction-topology signals into the attention mechanism, encouraging alignment between pore-level organization and atomic-scale environments. The transformer layers allow these representations to inform one another: high-persistence topological features can reinforce or modulate local cues, and vice versa. This design enables ITT to model both local chemistry and long-range pore-network topology, making it highly suitable for data-efficient, multi-property prediction.

Furthermore, ITT adopts a two-stage training strategy. During self-supervised pretraining, the model is exposed to over 0.6 million unlabeled porous structures spanning MOFs, COFs, ZEOs, PPNs, and others. The model learns to recover masked structural-level embeddings from the context of elemental, atomic, and interaction-topological embeddings—thus enforcing consistency between topological and chemical views without requiring external labels or handcrafted features. In the fine-tuning stage, ITT is adapted to supervised benchmarks across a wide range of properties, including adsorption (e.g., Henry coefficients), transport (e.g., diffusivity), and stability (e.g., mechanical resilience). Across these tasks, ITT achieves strong predictive performance and cross-family generalization, outperforming conventional baselines.

Finally, the multilevel design of ITT supports interpretability. Attention maps reveal that high-persistence structural and interaction tokens often align with physically meaningful regions—such as known adsorption sites or bottleneck apertures—while atomic-level contributions remain traceable through the CGCNN branch. Taken together, PIH plays a central role in enabling ITT to reconcile local chemical signals with global and relational topological features, forming a coherent and versatile architecture for learning from complex porous materials. Looking forward, the integration of interpretable topological and chemical reasoning in ITT holds strong potential for guiding the design of functional porous materials, accelerating materials discovery, and enabling robust, scalable modeling of large and structurally diverse frameworks encountered in real-world applications.
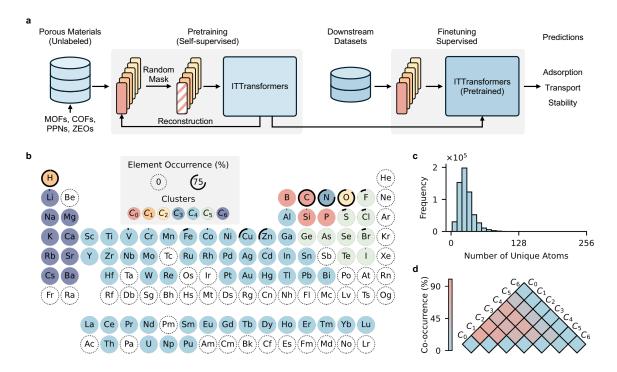
2.2 Schematic Workflow

Figure 2: Workflow and pre-analysis of the pretraining dataset. **a** Overall workflow: Unlabeled porous materials, including MOFs, COFs, PPNs, IZAs, are used for self-supervised pretraining, where random masking for structural embedding and reconstruction tasks are performed with ITT. The pretrained model is then fine-tuned on downstream datasets to predict adsorption, transport, and stability properties. **b** Elements in porous materials are grouped into seven clusters (C_0–C_6) by considering both co-occurrence statistics and chemical similarity. The size of each circle reflects element occurrence frequency, with H, C, N, and O dominating. **c** Unique atom distribution: Histogram of the number of unique atoms per structure across the pretraining dataset. **d** Pairwise co-occurrence percentages among the seven element clusters show strong correlations between different clusters but weak self-co-occurrence within the same cluster.

The schematic workflow of the proposed approach is illustrated in Figure 2**a**. To leverage the abundance of available structural data, we first collected a large and diverse data of porous material structures, without requiring any labeled property information, including MOFs, COFs, PPNs, and IZAs. In total, approximately 0.6 million structures were assembled for pretraining. Detailed sources and statistics for the collected data are provided in Table 2 and discussed in the Method section. Following data collection, we introduced the ITT model (Figure 1) to perform self-supervised pretraining. Each structure is first processed to generate multi-level embeddings. During pretraining, a subset of the structure-level embeddings is randomly masked, and the model is trained to reconstruct the masked portion based on the remaining multi-level context. This self-supervised objective requires no external labels, in contrast to previous supervised methods, and motivates the model to learn the underlying relationships of multi-level patterns across materials. By learning from intrinsic structural organization alone, the pretrained ITT becomes capable of modeling previously unseen porous materials [25]. In the fine-tuning stage, the same embedding procedure is applied to structures in labeled downstream datasets, and the pretrained ITT is adapted via supervised learning to predict specific properties. These downstream tasks span adsorption, trans-

port, and stability domains. The two-stage learning paradigm, which is self-supervised pretraining followed by task-specific fine-tuning, allows ITT to generalize effectively even in low-data regimes, while capturing both local and global structural signals critical to property prediction.

2.3 Data Preprocessing Rationale

Porous materials, especially MOFs, exhibit high elemental diversity, spanning over 70 elements in the periodic table. While the global dataset includes a vast variety of elements, each individual structure typically contains only a limited subset, which often just a few unique atomic environments. This imbalance poses a significant challenge for machine learning models: representing each element independently leads to sparse, high-dimensional inputs that are both inefficient and difficult to generalize from. Therefore, it is essential to develop a compact and chemically meaningful element representation that captures relevant distinctions without unnecessary complexity.

To address this, we performed a pre-analysis of element usage across approximately 0.6 million porous structures, with the goal of clustering elements into a small number of groups that preserve both chemical relevance and statistical interpretability. This clustering not only reduces input dimensionality but also enables robust modeling of inter-cluster interactions, which are central to our proposed ITT framework. The clustering strategy integrates two principles: (1) Co-occurrence statistics. Elements that frequently appear together are likely functionally distinct and should be separated to preserve interaction information. (2) Chemical similarity. Elements with similar properties are more likely to play analogous roles across materials and can be grouped accordingly. Specifically, to avoid biasing the final clustering results due to the data collection process, the verified and diverse ACR-MOF [33] dataset was used to determine the choice of clusters.

The strategy begins by filtering out rarely used or chemically unstable elements (atomic number > 103). A co-occurrence matrix is constructed to capture how often each element pair appears across structures. To avoid clustering highly co-occurring elements (which may be functionally complementary), we invert and normalize this matrix to serve as a dissimilarity signal. In parallel, we extract chemically meaningful features from the periodic table, i.e., atomic number, group, block, electronegativity, valence electrons, and atomic radius, and compute a cosine similarity matrix representing fundamental chemical likenesses. Next, we define anchor elements for initial cluster centers based on frequency and chemical diversity: C, H, O, N, and Zn, which dominate the dataset (top 5 leading elements) and represent core chemical functionalities. The remaining elements are grouped using a hybrid optimization of co-occurrence dissimilarity and chemical similarity, resulting in seven element clusters (C_0– C_6). The resulting clustering is visualized in Figure 2**b**, where black circle length reflects element frequency. As expected, H, C, N, and O dominate across the dataset, motivating their selection as separate clusters. The total count of elements along with their occurrence rates is shown in Supplementary Table S4. Additionally, the distribution of unique atomic environments per structure, shown in Figure 2**c**, demonstrates that most structures contain far fewer than 256 unique atoms, thereby justifying the choice of 256 as the maximum token number for atomic-level embeddings.

Figure 2**d** presents the pairwise co-occurrence heatmap between the seven clusters. The high co-occurrence between different clusters and low co-occurrence within the same cluster across all structures in the pretraining dataset supports the validity of the clustering strategy. Moreover, in 93.67% of the structures, the elements are distinctly assigned to different clusters. This separation

8

ensures that cluster-based abstraction does not obscure important interaction signals and element-level distinctions are preserved at the structural level, enabling accurate interaction modeling in the ITT framework.
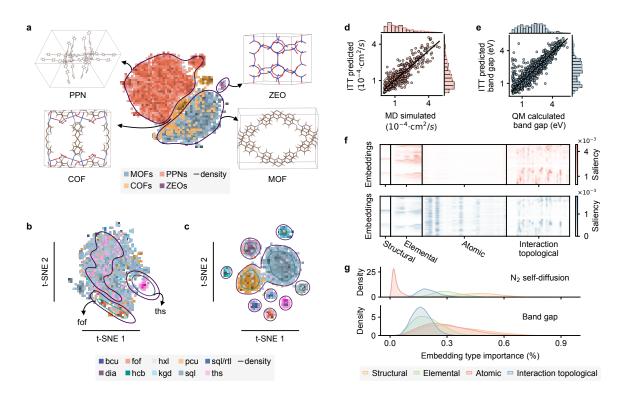
2.4 Results and Analysis of the ITT

Figure 3: ITT model analysis and visualization of representation learning. **a** 2D t-SNE visualization of the latent features (first token embeddings using for prediction) from the pretrained ITT model across four porous material types—MOFs, COFs, PPNs, and ZEOs—colored by class. Density contours (darker regions) indicate higher local concentration of samples, and representative structures are shown for each class. **b** t-SNE visualization of pretrained ITT's latent features for the CoRE MOF 2019 dataset, colored by topological class. Highlighted regions show dense clusters for specific topologies, such as ths and fof. **c** t-SNE visualization of the finetuned ITT model trained on a multi-class topology classification task using CoRE MOF 2019, showing increased separation and class-specific clustering of topologies. **d** Regression comparison between ITT-predicted and MD-simulated self-diffusion coefficients for N_2 (MAE= 3.23×10^{-5}), in units of 10^{-4} cm^2/s. Each dot represents one porous structure. **e** Comparison between ITT-predicted and QM-calculated band gaps (MAE=0.263 eV), where each dot corresponds to a single material. **f** Saliency score heatmaps for ITT models finetuned on N_2 self-diffusion (top) and band gap prediction (bottom), showing the contribution of each embedding token to the final prediction. Token groups include structural ($\times 1$), elemental ($\times 7$), atomic (≤ 256), and interaction topological embeddings ($\times 42$). The MOF (ID: 'ABAXUZ_FSR') from CoRE MOF 2019 dataset was used to generate the saliency score. (g) Distributions of embedding type importance across samples, shown for models finetuned on N_2 diffusion (top) and band gap (bottom). The x-axis represents the percentage contribution of each embedding type to the prediction, and the y-axis shows the sample density.

To assess the representational capacity of the pretrained ITT model, we visualize the learned latent features using 2D t-SNE projections. As shown in Figure 3**a**, the embeddings from the first token (prior to the prediction head) exhibit clear clustering by material type, despite being trained without labels. Distinct groupings are observed for well-refined datasets such as MOFs (CoRE

MOF 2019 [34]), COFs (CoRE-COFs v7 [35]), PPNs [36], and ZEOs (IZA-SC [37]), suggesting that ITT's self-supervised pretraining is capable of capturing underlying structural and chemical regularities across diverse porous materials. In contrast, when applied to more compositionally and structurally diverse hypothetical datasets, such as hPPNs [38] and MC-COFs [39], the resulting t-SNE embeddings (Supplementary Figure S1) display considerable overlap, indicating that additional supervision may be required to resolve fine-grained distinctions in complex systems. Beyond material-type separation, the pretrained ITT model also encodes topological organization within a single dataset (MOFs). Figure **3b** shows that structures in the CoRE MOF 2019 dataset cluster according to known topologies (e.g., fof, ths), despite no access to topology labels during training. This demonstrates the model's ability to learn meaningful topological features directly from structure. Upon finetuning ITT on a supervised classification task involving the 10 most frequent topologies, the latent space becomes more structured and class-discriminative (Figure **3c**), confirming the model's adaptability and effectiveness in downstream tasks.

Table 1: Performance comparison across downstream porous materials datasets[a]

Data Type	Dataset	Task	Size	Metric	Source	Scratch	ITT
MOFs	O_2/N_2-Select	Henry constant (N_2)	4744	R^2	0.70 [30]	0.80	**0.82**
		Henry constant (O_2)	5036	R^2	0.74 [30]	0.85	**0.86**
		Uptake (N_2)	5132	R^2	0.71 [30]	0.83	**0.83**
		Uptake (O_2)	5241	R^2	0.74 [30]	0.85	**0.87**
		Self-diffusivity (N_2, 1 bar)	5056	R^2	0.76 [30]	0.77	**0.81**
		Self-diffusivity (N_2, ∞ dilution)	5192	R^2	0.76 [30]	0.78	**0.82**
		Self-diffusivity (O_2, 1 bar)	5223	R^2	0.74 [30]	0.71	**0.76**
		Self-diffusivity (O_2, ∞ dilution)	5097	R^2	0.78 [30]	0.75	**0.79**
	CO_2–Henry	CO_2 Henry (log k_H)	9525	MAE	4.94e−01[b]	3.38e−01	**3.20e−01**
	QMOF–BD	Bandgap	20375	MAE	2.69e−01[b]	**2.63e−01**	2.72e−01
	SRS[c]	Solvent-removal stability	2179	Accuracy	**0.76**[d] [32]	0.71[d] (0.74)	0.70[d] (0.77)
	TST[c]	Thermal stability	3132	MAE	44[32]	46[d](50.2)	**43**[d](48.8)
COFs	MC–COF	CH_4 uptake (high P)	69840	MAE	2.17e+01[b]	2.95e+00	**2.35e+00**
		CH_4 uptake (low P)	69840	MAE	6.63e+00[b]	1.20e+00	**9.40e−01**
PPNs	hPPN	CH_4 uptake (1 bar)	17846	MAE	1.23e+00[b]	7.21e−01	**4.55e−01**
		CH_4 uptake (65 bar)	17846	MAE	7.70e+00[b]	3.52e+00	**3.09e+00**
ZEOs	ZEO–H	Henry constant (low P)	216	MAE	2.87e−01[b]	2.26e−01	**2.18e−01**
		Max loading (403 bar)	216	MAE	4.75e+00[b]	2.13e+00	**2.19e+00**

[a] Bolded values indicate the best performance for each corresponding dataset or task. [b] For datasets without reported source results, the CGCNN model is used as the baseline, with settings provided in Supplementary Information Section S3.2. [c] The text-mining labels were applied in the source paper[32]. [d] The predefined train/validation/test splits from the source paper were used [32].

We conducted a comprehensive evaluation of ITT on 18 downstream tasks covering adsorption, transport, stability, and electronic properties across MOFs, COFs, PPNs, and zeolites. When repeatable prior ML results were unavailable, we used CGCNN as a baseline under identical settings (details in Supplementary Information S3.2). For regression tasks we report MAE, RMSE, and R^2, and for classification we report Accuracy, Precision, and Recall. Aggregate outcomes are summarized in Table 1 (full metrics in Supplementary Information Table S6).

Against prior work and baselines, the proposed ITT achieves state-of-the-art results on 17 of 18 tasks under the same evaluation protocol. The only nuanced cases involve labels mined from the literature (SRS and TST), which can embed non-structural signals that structure-only models

do not consistently capture. Even so, ITT matches or exceeds prior reports in most settings: for example, ITT improves thermal-stability MAE on the source fixed split, while on SRS it is slightly below the reported fixed-split accuracy but surpasses the scratch baseline on the randomized split (Accuracy = 0.77).

Relative to the ITT scratch variant (no pretraining), self-supervised pretraining delivers consistent gains across medium/large datasets—particularly for adsorption and transport (all eight O_2/N_2 tasks, MC-COF and hPPN methane uptake, CO_2 Henry). Two small gaps remain: ZEO–H max loading (403 bar), which is a small (200) sample set where estimates are naturally high variance, shows virtually tied performance (MAE_{ITT}=2.19 vs. $MAE_{scratch}$=2.13), and QMOF-BD bandgap is marginally better with scratch in our runs. This behavior is consistent with the physics: bandgap is dominated by localized electronic environments, whereas our pretraining emphasizes global structural regularities; nevertheless, ITT's multilevel design (atomic, elemental, interaction-topological, structural) retains strong local cues and remains highly competitive.

Representative scatter plots for N_2 self-diffusion at 1 bar and bandgap (Figure **3d,e**) show tight alignment with ground truth (MD/DFT), indicating good calibration across both physical (transport) and electronic properties. Token-level saliency (Figure **3f**) provides a mechanistic view: for a representative MOF (ID: `ABAXUZ_FSR`), atomic- ($\sim$3 2%) and elemental- ($\sim$ 33%) level embeddings dominate bandgap prediction (blue), with interaction-topological ($\sim$ 19%) features next—consistent with local electronic structure; in contrast, transport relies more on structural/elemental/interaction-topological signals that encode pathways at larger scales (red). This pattern is further confirmed by the distribution of token type importance across dataset samples (Figure **3g**): band gap prediction exhibits high reliance on localized features (followed by elemental-level, interaction topological embeddings), while transport-related tasks favor global or intermediate-scale representations. It is also noticed at Figure **3d** that a small number of obvious outliers in the bandgap scatter dominate the residual tail, it motivates us report all test samples under the prescribed splits and all downstream datasets are provided and also the testing set ids. Together, ITT's self-supervised pretraining improves accuracy broadly, its multilevel embeddings adapt across disparate property types, and its saliency analysis yields interpretable attributions that align with known physics—local cues for electronic properties and broader structural cues for transport/adsorption.

3   Discussion

To better capture the structural and chemical complexity of porous materials, the proposed ITT framework incorporates the interaction topology, e.g., PIH. As illustrated in Figure 4, PIH generalizes classical persistent homology (PH) by introducing interaction-aware topological analysis, enabling the extraction of both global and local structural features across multiple atomic subsystems. Traditional graph-based representations have been widely used in materials modeling, yet they are fundamentally limited to pairwise atomic interactions and lack the expressiveness to capture higher-order relationships. In contrast, simplicial complexes extend beyond pairwise interactions by using higher-dimensional building blocks, i.e., 0-, 1-, and 2-simplices, to represent atoms, bonds, and triangular relationships, respectively (Figure **3a**). Classical PH, when applied to such complexes, captures the overall topology of entire structures (e.g., connectivity, loops,
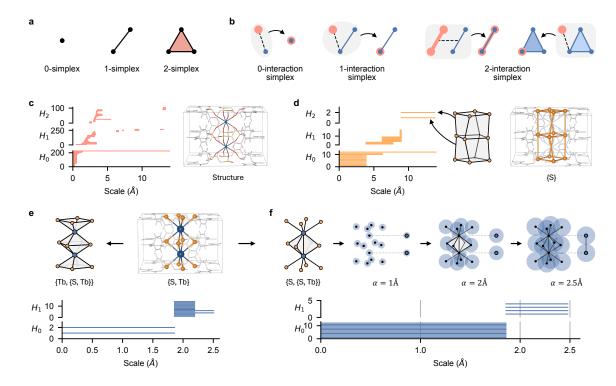
Figure 4: Interaction topology-based analysis. **a** Building blocks of a simplicial complex, including 0-simplex (vertex), 1-simplex (edge), and 2-simplex (filled triangle). **b** Building blocks of the interaction complex (n = 2), which generalizes simplicial complexes to describe interactions between two systems. 0-interaction simplex: intersection of two 0-simplices (vertices); 1-interaction simplex: intersection of a 0-simplex and a 1-simplex; 2-interaction simplices: formed either by the interaction of two 1-simplices or by the intersection of a 0-simplex with a 2-simplex. Pink and blue colors denote simplices from different systems. **c** Persistent homology (PIH, $n = 1$) in dimensions 0 to 2 for a MOF structure (ID: ABAXUZ_FSR), capturing the global topological features of the full structure. **d** Cluster-specific persistent homology (PIH, $n = 1$) barcodes for the S atoms in the same MOF structure. In dimension 2 (H_2), two prominent bars represent two cube-like cavities defined by S atoms arranged in octameric rings. **e** PIH ($n = 2$) refers to the interaction between the cluster containing S atoms and the cluster containing Tb atoms, where Tb is the interaction center. H_0 captures the number of Tb atoms and their local connectivity with nearby S atoms, whereas H_1 captures loop-like triangular structures around Tb formed within the interaction complex. **f** PIH ($n = 2$) results for the S-Tb interaction, centered at S. Three alpha interaction complexes at increasing filtration radii ($\alpha = 1$ Å, 2 Å, 2.5 Å) illustrate the multiscale construction of interaction features. The persistent H_1 bars reflect triangle-like loops formed between S and neighboring Tb atoms, while H_0 bars capture the number of distinct S centers contributing to the interaction complex.

voids), as demonstrated in Figure 4**c**. However, its focus on the global structure often overlooks critical element-specific and interaction-level details necessary for understanding chemically heterogeneous systems. To incorporate chemical specificity, PIH ($n = 1$) can be applied to elemental subsets of atoms. For example, Figure 4**d** shows the persistent barcodes derived from sulfur (S) atoms in a MOF structure, where two prominent H_2 bars represent cube-like cavities bounded by octameric S rings. While this provides localized chemical insight, treating each element subset independently may still fail to capture inter-element interactions, which are especially important in multi-component systems like MOFs and COFs. Similarly, emerging classes of materials such as POM-MOFs, which feature cavity-centered POM units embedded in larger frameworks [10], repre-

sent ideal candidates for PIH analysis due to the interplay between their internal cluster topology and the surrounding MOF network.

To address this, we introduce the concept of interaction complexes (ICs), shown in Figure 3**b**, which serve as the foundation for PIH with $n = 2$. This approach considers interactions between pairs of atomic groups by constructing joint simplicial structures based on their geometric overlap. For example, Figure 4**e, f** demonstrates how interactions between S and Tb atoms can be modeled from two perspectives—each using one element as the interaction center. When Tb is designated as the center (Figure 4**e**), H_0 captures the number and connectivity of Tb atoms with nearby S atoms, while H_1 reflects loop-like motifs formed in the vicinity of Tb. Conversely, centering on S atoms (Figure 4**f**) reveals different patterns: H_0 tracks distinct S-centered interaction sites, and H_1 identifies triangular loops involving neighboring Tb atoms. The construction of interaction complexes at multiple filtration scales (e.g., $\alpha = 1, 2, 2.5$ Å) highlights the multiscale nature of PIH, which can systematically capture geometric and topological transitions across spatial resolutions. In the context of POM-MOFs, this multiscale capability could be especially valuable for resolving cavity-specific interactions between POM clusters and the surrounding MOF scaffold, enabling a detailed topological characterization of these hybrid materials.

PIH's strength lies in its ability to integrate interaction-centered perspectives with topological analysis, preserving both fine-grained local interactions and larger-scale structural organization. Unlike traditional PH, which is structure-centric, PIH enables analysis of inter-system relationships, a critical capability for modeling functional behavior in chemically diverse materials. This dual-level characterization—distinguishing not only the topology within clusters but also the interactions between them—makes PIH a powerful foundation for the multi-level embeddings used in ITT. Ultimately, PIH enriches the ITT framework by providing topological features that encode both compositional and relational information, allowing the model to reason over the structural hierarchy inherent in real-world material systems.

Looking forward, the ITT framework holds particular promise for advancing the design and discovery of POM-based materials, such as POM-MOFs and polymer-modified POMs. These systems are especially interesting because POMs act as cavity-centered clusters whose functional properties, such as redox activity and charge states, are strongly modulated by their chemical surroundings (e.g., ligands, ordered MOF scaffolds, or polymer matrices) [8]. The multiscale capability of ITT is well suited for this challenge: atomic-level embeddings can capture charge-sensitive features, element- and pairwise-element embeddings from interaction topology can describe POM–ligand or POM–scaffold interactions, and structure-level embeddings can resolve the influence of long-range crystalline environments. Together, these representations could guide the rational design of POM-based materials with tailored electronic structures and memristive states, offering a systematic pathway for extending ITT to functional materials beyond conventional porous frameworks.

4 Method

4.1 Data Collection

In this work, we assembled a multi-domain corpus of porous materials, including MOFs, COFs, PPNs, and zeolites (ZEOs), to support structure-only self-supervised pretraining and supervised fine-tuning of the ITT model. For pretraining, we used five complementary structure collections

strictly as unlabeled crystals: ARC-MOF [33], Materials Cloud COFs (MC-COF) [39], CoRE-COF (v7) [35], amorphous PPNs [36], and IZA–SC zeolites [37]. ARC-MOF aggregates experimentally characterized and hypothetical MOFs; MC-COF comprises in-silico COFs built from a large linker library and relaxed for simulation; CoRE-COF (v7) provides curated experimental COF structures; the PPN set contains amorphous polymer networks generated via self-assembly simulations; and IZA–SC hosts canonical zeolite framework structures. In total, the pretraining pool includes 602,396 structures and can be readily expanded with large, label-free structure repositories. For fine-tuning, we assembled labeled datasets across MOFs, COFs, PPNs, and zeolites, using an 8:1:1 train/validation/test random split for every task. The MOF suite includes the eight-task N_2/O_2 selectivity collection derived from the simulated CoRE MOF 2019 database [34] (see Supplemental Table S7); to improve robustness, we removed extreme outliers using the upper-limit thresholds introduced by Orhan et al. [30], yielding a more uniform target-variable distribution. Additional MOF labels include Solvent-Removal Stability (binary activation stability) and Thermal Stability (decomposition temperatures) mined from the literature via text and image analysis [32], QMOF band gaps computed by periodic DFT workflows [16], and a CO_2 Henry-coefficient dataset reported as $\log k_H$ [23]. For COFs, we used the MC-COF dataset with methane-uptake labels generated by grand canonical Monte Carlo simulations [39]; for PPNs, the hPPN table with simulated methane-adsorption labels [38]; and for zeolites, the Zeolite_Hydrogen (ZEO–H) subset with labels (low-pressure slope and high-pressure loading) produced by high-throughput Monte Carlo hydrogen-adsorption simulations [40]. To ensure fair and comparable evaluation, we did not introduce any additional designed structures that could bias the model during either pretraining or fine-tuning; all datasets were collected from publicly available, fully traceable sources. Details on dataset sizes and brief descriptions are summarized in Table 2.

Table 2: Summary of porous materials datasets.

Purpose	Dataset	Size	Properties Used
Pretraining	ARC-MOF [33]	520,835	MOF structures only
	MC-COF [39]	69,840	COF structures only (Materials Cloud)
	CoRE-COF [35]	1,242	COF structures only (CoRE-COF version 7)
	PPN [36]	10,237	Amorphous PPN structures only
	IZA-SC [37]	242	Zeolite structures only
Downstream (MOFs)	O_2/N_2-Select [34, 30]	4,744–5,241	Henry constants, uptakes, and self-diffusivities for O_2/N_2 (8 tasks; see Table S7)
	SRS [32]	2,179	Solvent-removal stability (binary classification)
	TST [32]	3,132	Thermal stability temperature (K)
	QMOF-BD [16]	20,375	Bandgap (eV)
	CO_2–Henry [23]	9,525	CO_2 Henry constant ($\log k_H$; k_H in mol·kg^{-1}·Pa^{-1})
Downstream (COFs)	MC-COF [39]	69,840	CH_4 uptake at 65 bar and 5.8 bar (cm^3(STP)·cm^{-3})
Downstream (PPNs)	hPPN [38]	17,846	CH_4 uptake at 1 bar and 65 bar (cm^3(STP)·cm^{-3})
Downstream (ZEOs)	ZEO–H [40]	215	Henry constant at low P ((g/L)/bar), max loading at 403 bar (g/L)

4.2 Interaction Topology Theory

4.2.1 Interaction complex

To incorporate element-specific information when analyzing porous materials, we partition atoms into elemental clusters and study how these clusters jointly form topological features. At a fixed scale, let K be the simplicial complex on all atoms, and let $\{K_i\}_{i=1}^n$ be a family of subcomplexes of K induced by chosen element clusters (e.g., metals, oxygens, carbons). An interaction simplex is a tuple $(\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \in K_i$ whose intersection is nonempty,

$$\sigma_1 \cap \cdots \cap \sigma_n \neq \varnothing,$$

representing a joint feature contributed by all selected clusters. The demonstrations for 0-, 1-, and 2-interaction simplex are shown in Figure 4b. The resulting n-interaction complex is

$$\mathcal{K} = \big(K, \{K_i\}_{i=1}^n\big).$$

Specifically, when $n = 1$, $\mathcal{K} = (K, \{K\})$ reduces to the ordinary simplicial complex K. We use this in two ways: (i) structural-level embedding on all atoms (treating atoms as identical), and (ii) elemental-level embedding by applying the same construction to a single cluster K_i (capturing element-specific structure). When $n = 2$, $\mathcal{K} = (K, \{K_1, K_2\})$ captures pairwise interactions between two element clusters. For convenience, we write $\{K_i\}$ in place of $\big(K, \{K_i\}_{i=1}^n\big)$ whenever no confusion arises.

4.2.2 Interaction homology and its persistence

Analogous to ordinary homology, we form interaction chains by taking formal sums of interaction simplices and use a boundary operation that acts on one component at a time (the usual simplicial boundary applied componentwise while keeping the other components fixed). This identifies cycles (joint features with no boundary) and boundaries (those arising from higher-dimensional joint features). The resulting interaction homology group measures p-dimensional joint features—shared connected components ($p = 0$), co-supported loops/tunnels ($p = 1$), and co-supported cavities ($p = 2$) that exist because specific element clusters arrange together. The rank of the interaction homology group is the interaction Betti number

$$\beta_p = \dim_{\mathbb{K}} H_p(\{K_i\}; \mathbb{K}).$$

Here, $\mathbb{K}$ is the base field, and $H_p(\{K_i\}; \mathbb{K})$ denotes the p-dimensional interaction homology of $\{K_i\}$. To capture how these joint features change with spatial scale, we use persistent interaction homology. As the scale parameter a increases, both the whole complex $K(a)$ and each subcomplex $K_i(a)$ grow, producing a nested family of interaction complexes $\{K_i(a)\}$. For real numbers $a \leq b$, the inclusion $\{K_i(a)\} \to \{K_i(b)\}$ induced by $a \to b$ gives a map on the interaction homology. The (a, b)-persistent interaction homology group and its persistent Betti number are

$$H_p^{a,b}(\mathcal{F}_i; \mathbb{K}) = \mathrm{im}\Big(H_p\big(\mathcal{F}_i(a)\big) \to H_p\big(\mathcal{F}_i(b)\big)\Big), \qquad \beta_p^{a,b} = \dim_{\mathbb{K}} H_p^{a,b}(\mathcal{F}; \mathbb{K}),$$

where $\mathcal{F}(a) = (K(a), \{K_i(a)\}_{i=1}^n)$. Long-lived (persistent) joint features correspond to stable, chemically meaningful patterns, e.g., channels framed by metals and linkers—whereas short-lived

ones are often noise or highly localized artifacts. In this work, the alpha complex is used to represent a single system, while the interactions between alpha complexes are denoted by the alpha interaction complex (alpha-IC), as shown in Figure 4**f**.

If there is a single cluster ($n = 1$ and $K_1 = K$), the interaction homology reduces to the ordinary homology, and the persistent interaction homology reduces exactly to the classical persistent homology. In this work, the structural embedding corresponds to $n = 1$ on all atoms (Figure 4**c**); the elemental embedding corresponds to $n = 1$ on each cluster K_i (Figure 4**d**); and the interaction embedding uses $n = 2$ to quantify features arising specifically from the interaction of two element clusters, i.e., $H_p^{a,b}((K, \{K_1, K_2\}); \mathbb{K})$, providing element-aware, multiscale descriptors beyond what either cluster shows alone (Figure 4**e, f**). The detail and rigorous formulation of interaction topology are provided at the supplementary information Note 4.

4.3 Predictive Modeling

4.3.1 Data Preprocessing

In this work, each material's supercell is uniformly scaled to approximately $64\,\text{Å} \times 64\,\text{Å} \times 64\,\text{Å}$ to ensure consistent topological analysis across structures. For the structural- and elemental-level embeddings, we employ the HIP method with $n = 1$ (persistent homology) to compute topological invariants in the H_0, H_1, and H_2 homology dimensions. Those topological embeddings are generated using a grid-based approach: a grid from 0 to $25\,\text{Å}$ with a step size of $0.1\,\text{Å}$ is applied, and the Betti numbers—representing the number of topological features at each scale—are recorded. This produces a feature vector of length 750 (250 grid steps $\times$ 3 homology dimensions). The structural-level embedding consists of a single token (shape 1×750), while the elemental-level embedding contains seven tokens (7×750). Interaction-level topological embeddings are computed using the same grid for the H_0 and H_1 interaction homology dimensions, yielding a 500-dimensional feature vector for each elemental interaction pair. With seven elemental clusters, there are 42 distinct interaction pairs, resulting in 42×500 features in total. For atomic-level embeddings, duplicate atoms are removed by comparing local neighbor environments. Each remaining atom is initialized with features following the CGCNN scheme, using an $8\,\text{Å}$ cutoff to identify neighbors.

4.3.2 Computational settings

The IIT model consists of eight hidden layers (ITT decoders) with a hidden size of 256. The multi-head attention modules use eight heads, and the feed-forward layers have an intermediate size of 1024. For the atomic embeddings, four convolutional layers are applied, each producing atomic embeddings of size 128. During pretraining, the model is trained for 200 epochs with a batch size of 256 and the random mask rate 50% for the first token. For finetuning, a batch size of 32 is used for all downstream tasks, and training is run for 200 epochs. In both pretraining and finetuning, the AdamW optimizer is employed with a weight decay of 0.01 and a base learning rate of 0.0001. A warm-up learning rate schedule is applied, where the first 10% of the total epochs are designated as warm-up. Mean squared error is used as the loss function for regression tasks, and cross-entropy is used for classification tasks. To stabilize model performance, all labels are standardized before training to mitigate issues arising from extremely large value ranges. Additionally, for each downstream task, the predictions from the top five models (ranked by validation performance during

training) are averaged for each sample, resulting in more stable and robust outputs. Further details on the hyperparameters and model formulations are provided in Table S2-S4 and Supplementary Note 3.

5 Data availability

All porous-materials datasets analyzed in this study are publicly available from the sources cited in the manuscript. The curated bundle used in our studies, including strucature files and target labels, is trackable at `https://github.com/WeilabMSU/ITTransformer` to facilitate full reproducibility.

6 Code availability

The source code and model implementations used in this work are openly available at `https://github.com/WeilabMSU/ITTransformer`.

7 Acknowledgments

References

[1] Zhijie Chen, Kent O Kirlikovali, Karam B Idrees, Megan C Wasson, and Omar K Farha. Porous materials for hydrogen storage. *Chem*, 8(3):693–716, 2022.

[2] Rebecca L Siegelman, Eugene J Kim, and Jeffrey R Long. Porous materials for carbon dioxide separations. *Nature materials*, 20(8):1060–1072, 2021.

[3] Qian Li, Qun Li, Zhewei Wang, Xiaobo Zheng, Shichang Cai, and Jiabin Wu. Recent advances in hierarchical porous engineering of mofs and their derived materials for catalytic and battery: methods and application. *Small*, 20(8):2303473, 2024.

[4] Akashdeep Sharma, Sunil Babu Eadi, Hemanth Noothalapati, Michal Otyepka, Hi-Deok Lee, and Kolleboyina Jayaramulu. Porous materials as effective chemiresistive gas sensors. *Chemical Society Reviews*, 53(5):2530–2577, 2024.

[5] You Tao, Hui Liu, Hui-Yuan Kong, Xin-Yue Bian, Bin-Wei Yao, Yong Jun Li, Cheng Gu, Xuesong Ding, Lianfeng Sun, and Bao-Hang Han. Resistive memristors using robust electropolymerized porous organic polymer films as switchable materials. *Journal of the American Chemical Society*, 146(24):16511–16520, 2024.

[6] Yamil J Colón and Randall Q Snurr. High-throughput computational screening of metal–organic frameworks. *Chemical Society Reviews*, 43(16):5735–5749, 2014.

[7] Sangwon Lee, Baekjun Kim, Hyun Cho, Hooseung Lee, Sarah Yunmi Lee, Eun Seon Cho, and Jihan Kim. Computational screening of trillions of metal–organic frameworks for high-performance methane storage. *ACS Applied Materials & Interfaces*, 13(20):23647–23654, 2021.

[8] Hailong Yang, Guoli Yue, Tao Li, Lingling Du, Haohong Li, Zhirong Chen, and Shoutian Zheng. The old polyoxometalates in new application as molecular resistive switching memristors. *European Journal of Inorganic Chemistry*, 26(32):e202300264, 2023.

[9] Arash Ebrahimi, Lukáš Krivosudskỳ, Alexey Cherevan, and Dominik Eder. Polyoxometalate-based porphyrinic metal-organic frameworks as heterogeneous catalysts. *Coordination Chemistry Reviews*, 508:215764, 2024.

[10] Dong-Ying Du, Jun-Sheng Qin, Shun-Li Li, Zhong-Min Su, and Ya-Qian Lan. Recent advances in porous polyoxometalate-based metal–organic framework materials. *Chemical Society Reviews*, 43(13):4615–4632, 2014.

[11] Hilal Daglar and Seda Keskin. Recent advances, opportunities, and challenges in high-throughput computational screening of mofs for gas separations. *Coordination Chemistry Reviews*, 422:213470, 2020.

[12] Sakshi Yadav Schmid, Xiang Ma, Joshua A Hammons, Sebastian T Mergelsberg, Bradley S Harris, Thomas Ferron, Wenchao Yang, Wenhao Zhou, Renyu Zheng, Shuai Zhang, et al. Influence of peptoid sequence on the mechanisms and kinetics of 2d assembly. *ACS nano*, 18(4):3497–3508, 2024.

[13] Russell E Morris and Lee Brammer. Coordination change, lability and hemilability in metal–organic frameworks. *Chemical Society Reviews*, 46(17):5444–5462, 2017.

[14] Thomas D Bennett, François-Xavier Coudert, Stuart L James, and Andrew I Cooper. The changing state of porous materials. *Nature Materials*, 20(9):1179–1187, 2021.

[15] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: materials genomics and machine learning. *Chemical reviews*, 120 (16):8066–8129, 2020.

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

[17] Kaihang Shi, Zhao Li, Dylan M Anstine, Dai Tang, Coray M Colina, David S Sholl, J Ilja Siepmann, and Randall Q Snurr. Two-dimensional energy histograms as features for machine learning to predict adsorption in diverse nanoporous materials. *Journal of Chemical Theory and Computation*, 19(14):4568–4583, 2023.

[18] 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, 2018.

[19] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31 (9):3564–3572, 2019.

[20] Aditi S Krishnapriyan, Joseph Montoya, Maciej Haranczyk, Jens Hummelshøj, and Dmitriy Morozov. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Scientific reports*, 11(1): 8888, 2021.

[21] Dong Chen, Chun-Long Chen, and Guo-Wei Wei. Category-specific topological learning of metal–organic frameworks. *Journal of Materials Chemistry A*, 13(13):9292–9303, 2025.

[22] Dong Chen, Rui Wang, Guo-Wei Wei, and Feng Pan. Enhancing energy predictions in multi-atom systems with multiscale topological learning. *Journal of Materials Chemistry A*, 2025.

[23] 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, 2023.

[24] Eun Hyun Cho and Li-Chiang Lin. Nanoporous material recognition via 3d convolutional neural networks: Prediction of adsorption properties. *The journal of physical chemistry letters*, 12(9):2279–2285, 2021.

[25] 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.

[26] Jian Liu, Dong Chen, and Guo-Wei Wei. Persistent interaction topology in data analysis. *arXiv preprint arXiv:2404.11799*, 2024.

[27] Jian Liu, Dong Chen, and Guo-Wei Wei. Interaction homotopy and interaction homology. *arXiv preprint arXiv:2311.16322*, 2023.

[28] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.

[29] Debasis Banerjee, Cory M Simon, Anna M Plonka, Radha K Motkuri, Jian Liu, Xianyin Chen, Berend Smit, John B Parise, Maciej Haranczyk, and Praveen K Thallapally. Metal–organic framework with optimally selective xenon adsorption and separation. *Nature Communications*, 7(1):1–7, 2016.

[30] Ibrahim B Orhan, Hilal Daglar, Seda Keskin, Tu C Le, and Ravichandar Babarao. Prediction of o2/n2 selectivity in metal–organic frameworks via high-throughput computational screening and machine learning. *ACS Applied Materials & Interfaces*, 14(1):736–749, 2021.

[31] Hilal Daglar and Seda Keskin. Combining machine learning and molecular simulations to unlock gas separation potentials of mof membranes and mof/polymer mmms. *ACS Applied Materials & Interfaces*, 14(28):32134–32148, 2022.

[32] Aditya Nandy, Chenru Duan, and Heather J Kulik. Using machine learning and data mining to leverage community knowledge for the engineering of stable metal–organic frameworks. *Journal of the American Chemical Society*, 143(42):17535–17547, 2021.

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

[34] Yongchul G Chung, Emmanuel Haldoupis, Benjamin J Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S Camp, et al. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *Journal of Chemical & Engineering Data*, 64 (12):5985–5998, 2019.

[35] Minman Tong, Youshi Lan, Qingyuan Yang, and Chongli Zhong. Exploring the structure-property relationships of covalent organic frameworks for noble gas separations. *Chemical Engineering Science*, 168:456–464, 2017.

[36] Junkil Park, Wonseok Lee, and Jihan Kim. Large-scale construction and analysis of amorphous porous polymer network materials. *ACS Applied Materials & Interfaces*, 16(42):57190–57199, 2024.

[37] Ch Baerlocher, Lynne B McCusker, and David H Olson. *Atlas of zeolite framework types.* Elsevier, 2007.

[38] Richard L Martin, Cory M Simon, Berend Smit, and Maciej Haranczyk. In silico design of porous polymer networks: high-throughput screening for methane storage materials. *Journal of the American Chemical Society*, 136(13):5006–5022, 2014.

[39] Rocío Mercado, Rueih-Sheng Fu, Aliaksandr V Yakutovich, Leopold Talirz, Maciej Haranczyk, and Berend Smit. In silico design of 2d and 3d covalent organic frameworks for methane storage applications. *Chemistry of Materials*, 30(15):5069–5086, 2018.

[40] Yangzesheng Sun, Robert F DeJaco, Zhao Li, Dai Tang, Stephan Glante, David S Sholl, Coray M Colina, Randall Q Snurr, Matthias Thommes, Martin Hartmann, et al. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Science Advances*, 7(30):eabg3983, 2021.