Supplemental Information
Category-Specific Topological Learning of Metal-Organic Frameworks

Dong Chen[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*

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

Contents

S1 Evaluation metrics

In this work, the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 (Coefficient of Determination) were used for evaluating machine learning models.

The Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It quantifies the difference between the predicted values ($\hat{y}_i$) and the true values (y_i). The formula for RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{1}$$

where: 1) n is the number of observations. 2) $\hat{y}_i$ is the predicted value for the i-th observation. 3) y_i is the actual value for the i-th observation. 4) Lower values of RMSE indicate a better fit of the model to the data. It has the same unit as the response variable.

The Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. It calculates the average of the absolute differences between predicted and actual values. The formula for MAE is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

where: 1) n is the number of observations. 2) $\hat{y}_i$ is the predicted value for the i-th observation. 3) y_i is the actual value for the i-th observation. 4) MAE provides a straightforward interpretation of the average error. Like RMSE, lower values indicate a better fit, and it has the same unit as the response variable.

The Coefficient of Determination (R^2) measures the proportion of variance in the dependent variable that is predictable from the independent variables. It essentially indicates how well the model fits the data. The R^2 value ranges from 0 to 1, with 1 indicating a perfect fit. The formula for R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where: 1) n is the number of observations. 2) y_i is the actual value for the i-th observation. 3) $\hat{y}_i$ is the predicted value for the i-th observation. 4) $\bar{y}$ is the mean of the observed values. 5) An R^2 value close to 1 means that the model explains a large portion of the variance, whereas a value close to 0 indicates that the model explains very little variance.

S2 Datasets comparison

In this work, we compared the proposed method with the descriptor-based method [1], MOF-Transformer [2], and PMTransformer [3]. Although the eight datasets listed in Table 2 are derived from the study by Orhan et al. [1], their sizes differ slightly. The datasets used in this work were directly generated from the source repository (https://github.com/ibarisorhan/MOF-O2N2/tree/main/mofScripts) established by Orhan et al. [1]. While the datasets used in MOF-Transformer [2] and PMTransformer [3] were also obtained from the same source, their exact data details were not explicitly provided. To ensure a clear and transparent comparison, we compiled the data information for all methods compared in this study, as summarized in Table S1.

Table S1: Comparison of datasets used in the published works across various MOF datasets.

Datasets	**CSTL**	**Descriptor-based**[1]	**MOFTransformer**[2]	**PMTransformer**[3]
Henry's constant N_2	4744	4755		
Henry's constant O_2	5036	5045		
N_2 uptake (mol/kg)	5132	5158	5286	5286
O_2 uptake (mol/kg)	5241	5259	5286	5286
Self-diffusion of N_2 at 1 bar (cm^2/s)	5056	5079	5286	5286
Self-diffusion of N_2 at infinite dilution (cm^2/s)	5192	5202		
Self-diffusion of O_2 at 1 bar (cm^2/s)	5223	5247	5286	5286
Self-diffusion of O_2 at infinite dilution (cm^2/s)	5097	5115		

S3 Model Repeatability

To ensure robust evaluation, we repeated the random data split 10 times, and for each split, 10 models were trained with different random seeds, resulting in a total of 100 models per dataset. The performance metrics, including RMSE, MAE, and r^2 correlation, were averaged over these 100 models and reported as the final results. This approach of using a single set of hyperparameters and a consistent evaluation protocol highlights the robustness of the predictive model, making the results reliable and comparable to existing methods in the literature. Specifically, the Gradient Boosting Tree (GBT) model was constructed to perform regression analysis using the proposed category-specific topological learning (CSTL) embedding as input features. We implemented the gradient boosting regressor from Scikit-learn [4], optimizing the squared error loss function. The model parameters were set as follows: max_depth=7, max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=10,000, and subsample=0.5. The heatmap of MAE/r^2/RMSE values for 100 predictive models across eight datasets are shown in Figure S2, Figure S3, and Figure S4.

S4 Topological objects

Graph. The graph is a key structure for illustrating relationships among different entities, representing one of the most prevalent data forms. It is composed of nodes (or vertices) and edges, which establish the connections between these nodes. Graphs can be enhanced in several ways, such as by adding directionality to create directed graphs (digraph), assigning weights for weighted graphs, or incorporating geometric properties in geometric graphs. These enhanced graphs are excellent for representing relationships and attributes in various scenarios. Formally, a graph is defined as a pair (V, E), where V represents a set of vertices and E, a subset of $V \times V$, signifies the set of edges. Vertices and edges are the core components of a graph. Tools like adjacency matrices, degree matrices, and Laplacian matrices are utilized to describe the interactions between vertices and edges. These matrices are pivotal in graph theory and network analysis, capturing the graph's underlying topological structure. Although graphs are inherently one-dimensional, methods from simplicial complexes are sometimes used to express the graph's higher-dimensional aspects.

Simplicial complex. A simplicial complex is a type of topological space constructed from basic units known as simplices. A simplex extends the notion of a triangle or tetrahedron to any number of dimensions. For a set of vertices V, a k-simplex σ_k is typically represented by a subset of V containing $k+1$ elements, and is expressed as $\sigma = \langle v_0, v_1, \ldots, v_k \rangle$. Any subset of σ_{k-1} is considered a face of σ_k.

A *simplicial complex*, denoted as K, based on a vertex set V, is defined by a group of simplices that meet two criteria: (1) If a simplex σ is part of K, then all of its faces, including individual vertices, are also included in K; (2) The intersection of any two simplices within K is either empty or a face (subset) common to both simplices. From these characteristics, it's evident that a graph can be interpreted as a 1-dimensional simplicial complex, where its simplices consist of vertices (0-simplices) and edges (1-simplices).

In a k-simplex, the boundary is the set of its $(k-1)$-dimensional faces. The *boundary operator*, symbolized as ∂_k, operates on a k-simplex $\langle v_0, v_1, \ldots, v_k \rangle$ in the following mathematical form:

$$\partial_k \langle v_0, v_1, \ldots, v_k \rangle = \sum_{i=0}^{k} (-1)^i \langle v_0, \ldots, \widehat{v_i}, \ldots, v_k \rangle, \tag{4}$$

where $\widehat{v_i}$ indicates the exclusion of the vertex v_i. A chain complex is a series of Abelian groups (or modules), interconnected by boundary operators. Suppose G is an Abelian group. The k-th group in the chain complex, denoted as $C_k(K; G)$, comprises formal sums of k-simplices. The boundary operator $\partial_k : C_k(K; G) \to C_{k-1}(K; G)$ maps a k-simplex to its $(k-1)$-dimensional boundary. The sequence of the chain complex can be represented as:

$$\cdots \xrightarrow{\partial_{k+1}} C_k(K; G) \xrightarrow{\partial_k} C_{k-1}(K; G) \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_2} C_1(K; G) \xrightarrow{\partial_1} C_0(K; G). \tag{5}$$

A critical characteristic of the boundary operator is that the composition of two consecutive boundary operators equals zero, i.e., $\partial_{k-1} \circ \partial_k = 0$. This implies that the boundary of a boundary is always null, carrying significant topological implications. The structure of the chain complex provides a systematic way to analyze how boundaries integrate with each other. Beyond the simplicial

complex, other topological objects—such as the clique complex, cell complex, cellular sheaf [5], hypergraph, neighborhood complex [6, 7], Hom complex, knot, link, and tangle [8, 9]—can be further explored in the analysis of the given data.

S5 Supplementary tables

In the following section, we provide supplementary tables that offer additional data and insights pertinent to our study. Readers are encouraged to refer to these tables for a more detailed exploration of the topics covered in the main text.

Table S2: Comparison of CSTL models with different training-test splits.

Datasets	CSTL(80% training, 10% test)			CSTL(80% training, 20% test)		
	r^2	mae	rmse	r^2	mae	rmse
Henry's constant N_2	0.80	4.90E-07	7.25E-07	0.79	4.98E-07	7.36E-07
Henry's constant O_2	0.83	4.98E-07	7.63E-07	0.83	5.00E-07	7.69E-07
N2 uptake (mol/kg)	0.79	4.98E-02	7.37E-02	0.79	4.98E-02	7.39E-02
O2 uptake (mol/kg)	0.85	4.50E-02	6.82E-02	0.85	4.54E-02	6.90E-02
Self-diffusion of N_2 at 1 bar (cm2/s)	0.80	3.40E-05	4.69E-05	0.80	3.39E-05	4.64E-05
Self-diffusion of N_2 at infinite dilution (cm2/s)	0.80	3.75E-05	5.15E-05	0.80	3.79E-05	5.21E-05
Self-diffusion of O_2 at 1 bar (cm2/s)	0.82	3.21E-05	4.45E-05	0.81	3.32E-05	4.62E-05
Self-diffusion of O_2 at infinite dilution (cm2/s)	0.79	3.34E-05	4.53E-05	0.79	3.35E-05	4.54E-05

Table S3: Comparison of CSTL models (using only C_{all} features) with various training-test splits.

Datasets	CSTM(80% training, 10% test)			CSTM(80% training, 20% test)		
	r2	mae	rmse	r2	mae	rmse
Henry's constant N2	0.70	6.17E-07	8.75E-07	0.70	6.22E-07	8.81E-07
Henry's constant O2	0.74	6.52E-07	9.47E-07	0.74	6.54E-07	9.55E-07
N2 uptake (mol/kg)	0.71	6.14E-02	8.64E-02	0.71	6.18E-02	8.71E-02
O2 uptake (mol/kg)	0.77	5.90E-02	8.54E-02	0.76	6.04E-02	8.72E-02
Self-diffusion of N2 at 1 bar (cm2/s)	0.78	3.54E-05	4.82E-05	0.79	3.52E-05	4.77E-05
Self-diffusion of N2 at infinite dilution (cm2/s)	0.78	3.94E-05	5.38E-05	0.78	3.97E-05	5.41E-05
Self-diffusion of O2 at 1 bar (cm2/s)	0.80	3.41E-05	4.65E-05	0.79	3.51E-05	4.81E-05
Self-diffusion of O2 at infinite dilution (cm2/s)	0.78	3.51E-05	4.70E-05	0.77	3.52E-05	4.72E-05

S6 Supplementary figures

In this section, we present a series of supplementary figures that further elucidate and complement the findings discussed in the main text. Readers are encouraged to consult these figures for a richer understanding and visual representation of the concepts and results introduced in the main manuscript.

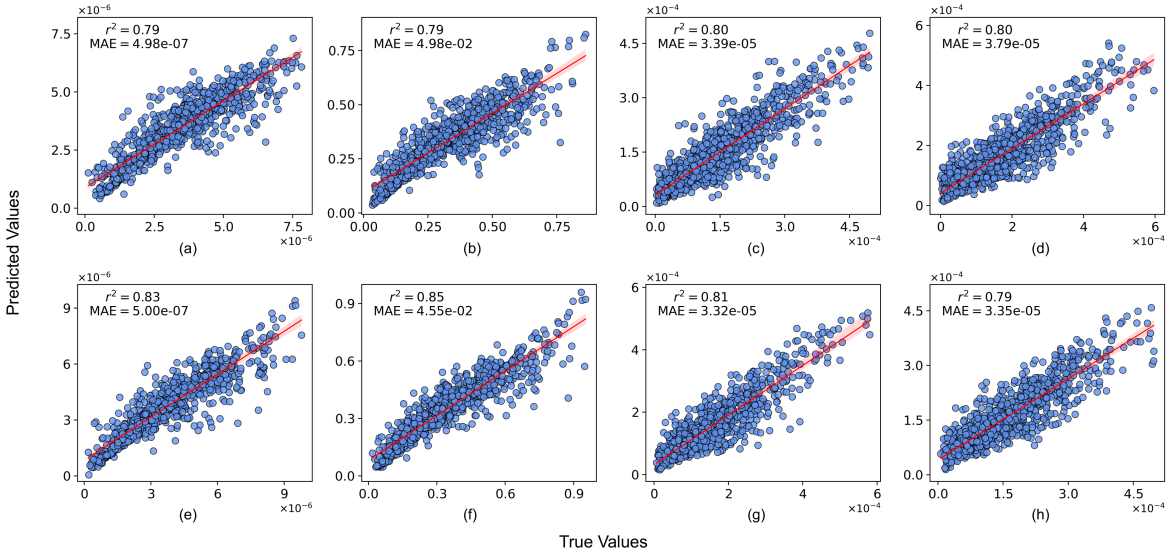

Figure S1: Comparison between predicted and true values for eight datasets on O_2/N_2 selectivity properties in MOF materials. Panels **a**-**h** show prediction performance for different properties: Henry's constant for N_2/O_2 (a, e), N_2/O_2 uptake (mol/kg) (b, f), self-diffusivity of N_2/O_2 at 1 bar (cm^2/s) (c, g), and self-diffusivity of N_2/O_2 at infinite dilution (cm^2/s) (d, h). Each panel displays the R^2 and the MAE in the upper left corner. Each dataset was randomly split, with 80% used for training and rest 20% for testing.

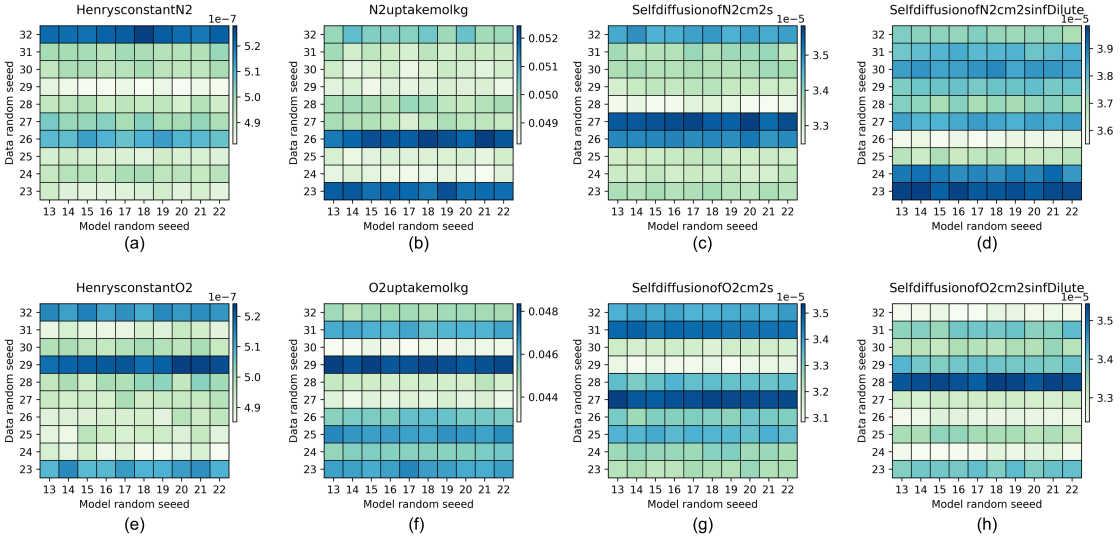

Figure S2: Heatmap of MAE values for predictive models across eight datasets related to O_2/N_2 selectivity properties in MOF materials. Panels (a)-(h) represent the MAE results for properties including Henry's constant for N_2 (a) and O_2 (e), N_2/O_2 uptake (mol/kg) for N_2 (b) and O_2 (f), self-diffusivity at 1 bar (cm^2/s) for N_2 (c) and O_2 (g), and self-diffusivity at infinite dilution (cm^2/s) for N_2 (d) and O_2 (h). Each dataset was randomly split 10 times with seeds ranging from 23 to 32, reserving 80% for training and 10% for testing. For each split, 10 separate models were trained with random seeds from 13 to 22, resulting in a total of 100 models per dataset. The heatmap color bar illustrates the MAE values for these 100 models, providing insight into prediction variability across different datasets and modeling scenarios.

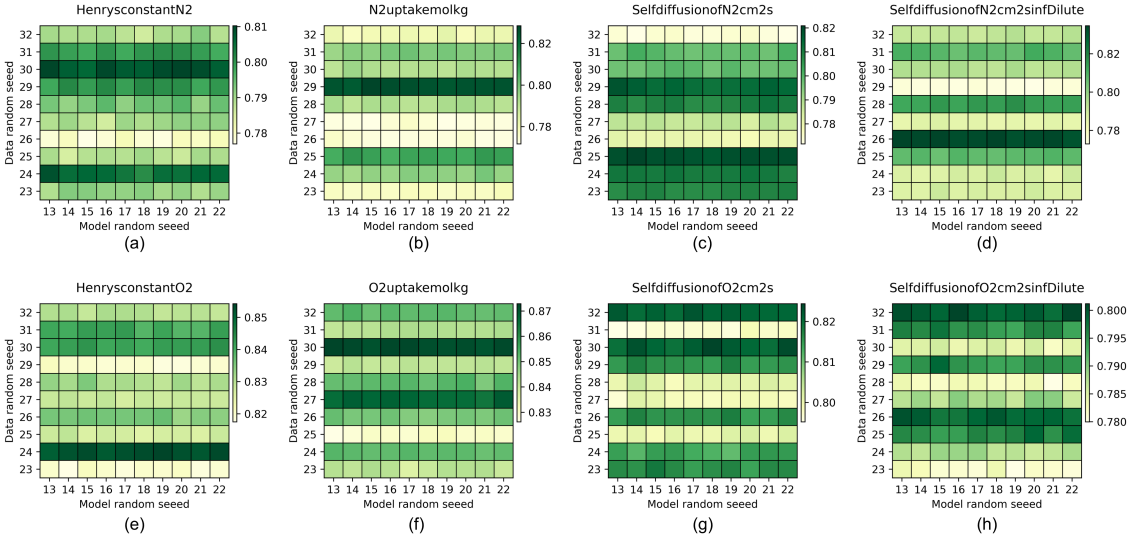

Figure S3: Heatmap of r^2 values for predictive models across eight datasets related to O_2/N_2 selectivity properties in MOF materials. Panels (a)-(h) represent the MAE results for properties including Henry's constant for N_2 (a) and O_2 (e), N_2/O_2 uptake (mol/kg) for N_2 (b) and O_2 (f), self-diffusivity at 1 bar (cm^2/s) for N_2 (c) and O_2 (g), and self-diffusivity at infinite dilution (cm^2/s) for N_2 (d) and O_2 (h). Each dataset was randomly split 10 times with seeds ranging from 23 to 32, reserving 80% for training and 10% for testing. For each split, 10 separate models were trained with random seeds from 13 to 22, resulting in a total of 100 models per dataset. The heatmap color bar illustrates the r^2 values for these 100 models.

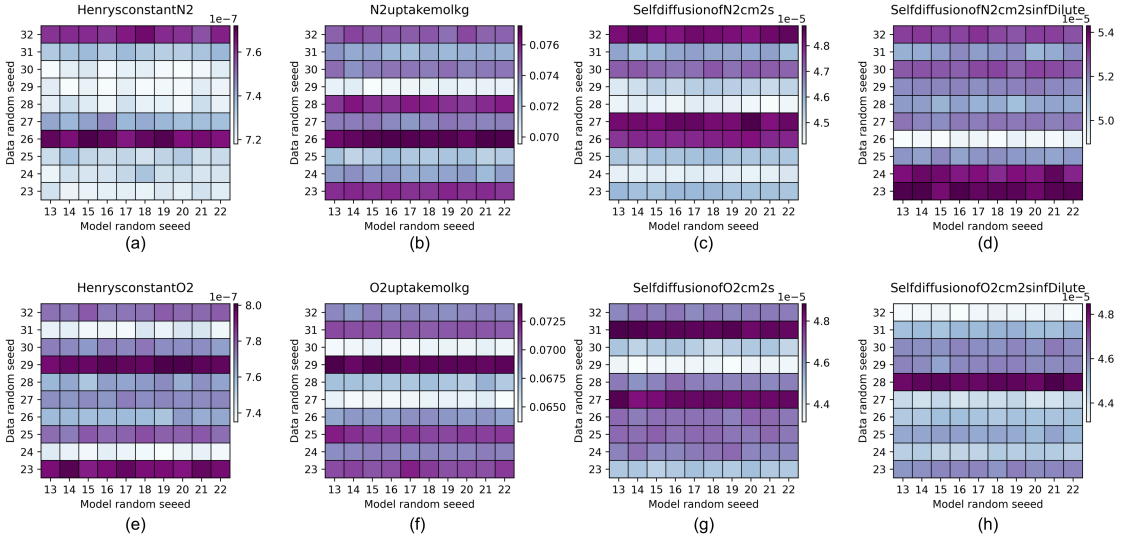

Figure S4: Heatmap of RMSE values for predictive models across eight datasets related to O_2/N_2 selectivity properties in MOF materials. Panels (a)-(h) represent the MAE results for properties including Henry's constant for N_2 (a) and O_2 (e), N_2/O_2 uptake (mol/kg) for N_2 (b) and O_2 (f), self-diffusivity at 1 bar (cm^2/s) for N_2 (c) and O_2 (g), and self-diffusivity at infinite dilution (cm^2/s) for N_2 (d) and O_2 (h). Each dataset was randomly split 10 times with seeds ranging from 23 to 32, reserving 80% for training and 10% for testing. For each split, 10 separate models were trained with random seeds from 13 to 22, resulting in a total of 100 models per dataset. The heatmap color bar illustrates the RMSE values for these 100 models.

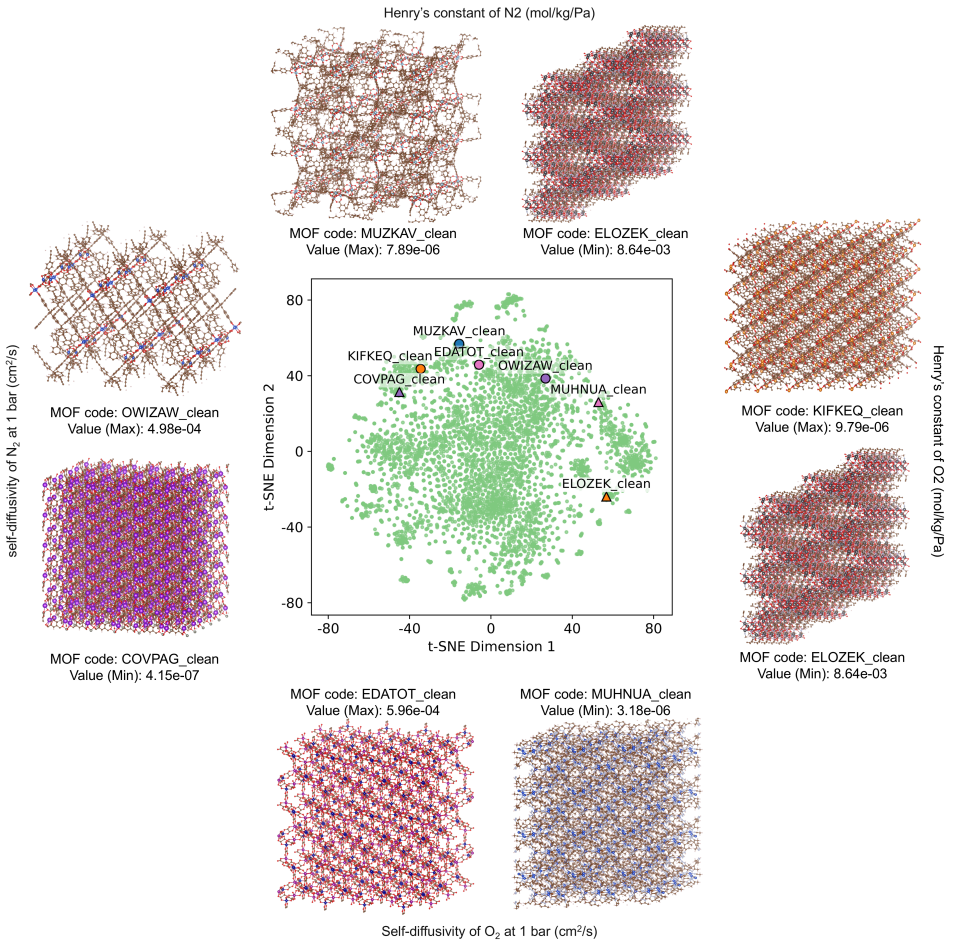

Figure S5: t-SNE feature reduction for category-specific topological features of MOF materials, where each green point represents a distinct MOF material. Highlighted circles and triangles indicate materials with maximum and minimum values, respectively, for four key properties: Henry's constant for N_2, Henry's constant for O_2, self-diffusivity of N_2 at 1 bar (cm^2/s), and self-diffusivity of O_2 at 1 bar (cm^2/s). 3D structures of the materials with minimum and maximum values for each property are shown around the t-SNE plot.

References

[1] 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.

[2] 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.

[3] 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.

[4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[5] Xiaoqi Wei and Guo-Wei Wei. Persistent topological laplacians–a survey. *arXiv preprint arXiv:2312.07563*, 2023.

[6] Jian Liu, Dong Chen, Feng Pan, and Jie Wu. Neighborhood path complex for the quantitative analysis of the structure and stability of carboranes. *Journal of Computational Biophysics and Chemistry*, 22(04):503–511, 2023.

[7] Jian Liu, Dong Chen, Jingyan Li, and Jie Wu. Neighborhood hypergraph model for topological data analysis. *Computational and Mathematical Biophysics*, 10(1):262–280, 2022.

[8] Dimitry Kozlov. *Combinatorial algebraic topology*, volume 21. Springer Science & Business Media, 2007.

[9] Li Shen, Jian Liu, and Guo-Wei Wei. Evolutionary khovanov homology. *AIMS Mathematics*, 9 (9):26139–26165, 2024.