Electronic Supplementary Information (ESI)

Graph Convolutional Network-Guided Inverse Link Prediction for Sparsification of Metal–Organic Framework Graphs in Large-Scale Cheminformatics

Elnaz Bangian, Mehrdad Jalali*, Mahboobeh Houshmand

Table of Contents

Section S1. Computational Complexity Comparison between ILP-based and Edge Betweenness Centrality (EBC)-based Sparsification

In Table S1, we present a detailed comparison between the ILP-based sparsification approach and the classical Edge Betweenness Centrality (EBC) method. The ILP approach involves multiple computational steps, including neural network training, feature-based edge importance prediction, and hyperparameter optimization, leading to a complexity of $O(p \cdot k \cdot E \cdot n_f)$, where p is the number of hyperparameter search steps, k is the number of training epochs, E is the number of edges, and n_f is the number of features. On the other hand, the EBC method relies purely on graph structure, calculating edge betweenness with a complexity of $O(nm)$and then removing high-centrality edges. Although both methods scale with the size of the graph, EBC is significantly more efficient for medium to large graphs due to the absence of learning

and hyperparameter tuning, making it more suitable for large-scale applications. Furthermore, EBC provides a more interpretable sparsification process based on the topological importance of edges, while ILP captures additional data-driven patterns at the cost of higher computational demand.

Table S1. Comparison of Computational Complexity and Characteristics of ILP and EBC Sparsification Methods

Step	ILP	EBC (Edge Betweenness Centrality)
Edge importance computation	Neural Network prediction for all edges	Betweenness Centrality computation
Edge importance complexity	$O(n_f \cdot E)$where n_f is the number of input features, E is the number of edges (per forward pass)	$O(n \cdot m)$ using Brandes' algorithm (exact computation)
Neural network training	$O(k \cdot E \cdot n_f)$ where k is the number of epochs	Not needed
Modularity computation	$O(m)$ for each community modularity calculation	Optional (used only if it measures quality, not needed for sparsification)
Hyperparameter Search (for best-a)	Typically, $O(p.(kEn_f + m)$ if p with different a values	Not needed
Total	$O(p.(kEn_f + m)$	$O(n \cdot m)$
Edge removal	$O(E)$	$O(Elog\ E)$sorting step for top edges)
Node removal (isolates)	$O(n)$	$O(n)$

Section S2. Inverse Link Prediction (ILP) Sparsification Analysis at Multiple Threshold Levels

Table S2 summarizes the network statistics obtained from applying ILP-based graph sparsification across a range of thresholds. The reported metrics include the number of remaining nodes and edges, node and edge reduction percentages, average degree, modularity, and the number of detected communities. The threshold refers to the cut-off applied to the ILP-derived edge weights during sparsification.

In this work, we selected three representative thresholds (0.90, 0.95, and 0.98) to systematically analyze the trade-off between network simplification and structural preservation. These thresholds represent distinct sparsification regimes, ranging from moderate to highly aggressive, allowing us to evaluate the flexibility and robustness of the proposed method.

- **Threshold = 0.90 (Moderate Sparsification):** The network experiences a substantial edge reduction (~90.75%) and moderate node reduction (~28.36%) while maintaining an average degree of ~8.53 and a modularity of 0.85. This level is suitable for scenarios where it is essential to retain a considerable portion of the network's connectivity and community structure.
- **Threshold = 0.95 (Strong Sparsification):** At this threshold, the network undergoes a more significant sparsification with ~96.97% edge reduction and ~47.78% node reduction. Despite the increased sparsification, modularity improves slightly to ~0.87, and the network still preserves a clear community structure. This stage is particularly useful for tasks focusing on community separation and reducing redundancy.

- **Threshold = 0.98 (Aggressive Sparsification):** In this highly aggressive setting, the network undergoes ~99.54% edge reduction and ~82.86% node reduction. Interestingly, the modularity continues to improve (up to ~0.94), emphasizing the contrast between communities even under severe sparsification. However, the average degree drops to ~1.78, leading to a highly simplified network, which may be favorable for visualization, pattern extraction, or interpretability tasks.

The trends associated with modularity improvement, node and edge reduction behavior, and average degree decay across all threshold levels are further illustrated in **Figure S2**, which shows how the network progressively evolves through the sparsification process. The figure demonstrates that while edge removal is highly effective, the ILP framework carefully preserves key nodes and community structures across different sparsification stages.

Table S2. Network statistics after ILP-based sparsification across multiple thresholds.

Threshold	Remaining Nodes	Remaining Edges	Node Reduction (%)	Edge Reduction (%)	Average Degree	Modularity	Communities
0.5	12016	246057	4.34	40.66	40.95	0.83	769
0.8	10622	90148	15.44	78.26	16.97	0.84	683
0.9	8999	38365	28.36	90.75	8.53	0.85	686
0.91	8719	33245	30.59	91.98	7.63	0.85	685
0.92	8360	28067	33.44	93.23	6.71	0.85	705
0.93	7938	22914	36.80	94.47	5.77	0.86	706
0.94	7414	17819	40.98	95.70	4.81	0.86	725
0.95	6559	12548	47.78	96.97	3.83	0.87	748
0.96	5179	7430	58.77	98.21	2.87	0.89	775
0.97	3400	3780	72.93	99.09	2.22	0.92	726
0.98	2153	1920	82.86	99.54	1.78	0.94	574
0.99	1238	885	90.14	99.79	1.43	0.95	410

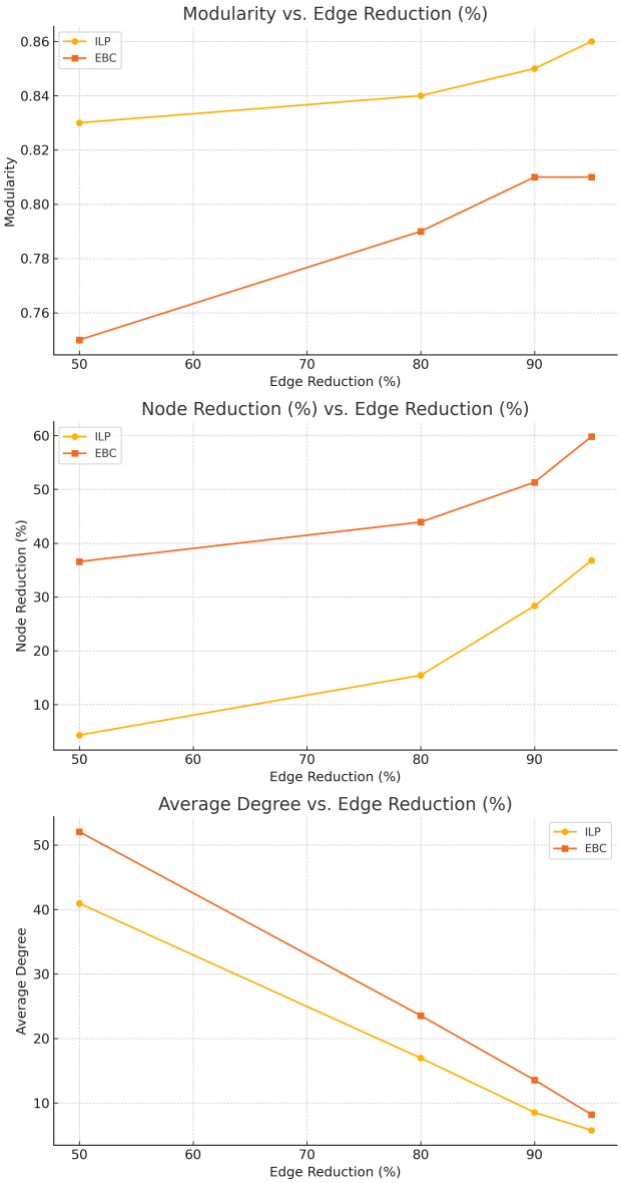

Figure S2. ILP outperforms EBC by preserving higher modularity, retaining more nodes, and achieving lower average degrees across all sparsification levels.

Section S3. Explanation of Variables Used in the ILP-GCN Sparsification Algorithm

All variables and notations used in the ILP-GCN sparsification algorithm are summarized in **Table S3**, which provides a detailed glossary explaining the meaning, type, and role of each variable involved in graph construction, link prediction, and sparsification. This table serves as a reference for understanding the pseudo-code and implementation details provided in Section 2.2.

Table S3. Glossary of Variables Used in the ILP-GCN Algorithm

Variable	Description	Data Type	Role
G	Input graph with nodes (V) and edges (E)	NetworkX Graph	Represents the MOF network before sparsification
V	Set of nodes	List / Set	Nodes correspond to MOFs
E	Set of edges	List / Set of tuples	Edges represent similarity links between MOFs

X	Node feature matrix	Float32 matrix of shape (n, f)	Contains features of nodes, including fingerprints, metal descriptors, and geometric features
ε	Stability constant	Float	Small positive constant to avoid division by zero during inverse calculations
α	Normalization factor	Float	Scales ILP-GCN-derived weights
γ	Balance factor	Float ($0 \leq \gamma \leq 1$)	Balances between initial graph weights and ILP-GCN-derived weights
Θ	GCN parameters	Tensor (trainable)	Includes all learnable parameters of the GCN (weight matrices)
S_GCN(e_ij)	GCN-based link prediction score	Float	Predicts the existence probability of edge (i, j)
I(e_ij)	ILP weight	Float	Inverse of GCN prediction score to identify removable edges
W_initial(e_ij)	Original edge weight	Float	Derived from similarity scores (linkers and metal descriptors)
W_ILP-GCN(e_ij)	ILP-GCN derived edge weight	Float	Weight based on inverse link prediction
W_final(e_ij)	Final edge weight	Float	Combined weight used for modularity optimization and sparsification
T	Threshold set	List of floats	Set of thresholds tested for sparsification (e.g., [0.90, 0.95, 0.98])
Removed_Edges	List of removed edges	List of tuples	Edges removed during sparsification

Section S4. Comparative Evaluation of ILP and EBC Sparsification

In this section, we provide a direct comparison between the proposed Inverse Link Prediction (ILP) sparsification and the classical Edge Betweenness Centrality (EBC) method. To ensure a fair evaluation, we matched both methods according to similar edge reduction ratios (50%, 80%, 90%, and 95%). **Table S4** summarizes the key structural metrics including node reduction, edge reduction, modularity, and average degree. The results clearly demonstrate that ILP consistently achieves higher modularity values while preserving more nodes compared to EBC, particularly at higher sparsification levels. Moreover, ILP results in a more gradual reduction of the average node degree, preserving more of the network's connectivity structure. These findings confirm ILP's advantage in maintaining the structural and community integrity of MOF networks under different levels of sparsification.

Table S4. Comparison of ILP and EBC Sparsification Results at Similar Edge Reduction Ratios

Removal Ratio	ILP Threshold	ILP Node Reduction (%)	EBC Node Reduction (%)	ILP Edge Reduction (%)	EBC Edge Reduction (%)	ILP Modularity	EBC Modularity	ILP Avg Degree	EBC Avg Degree
0.5	0.50	4.34	36.57	40.66	50	0.83	0.75	40.95	52.05
0.8	0.80	15.44	43.92	78.26	80	0.84	0.79	16.97	23.55
0.9	0.90	28.36	51.32	90.75	90	0.85	0.81	8.53	13.56
0.95	0.93	36.80	59.80	94.47	95	0.86	0.81	5.77	8.21

The comparative effects of ILP and EBC sparsification on key network properties, including modularity, node reduction, and average degree under varying edge reduction levels, are summarized in **Figure S4**, demonstrating the advantages of ILP in preserving network structure while achieving effective sparsification.

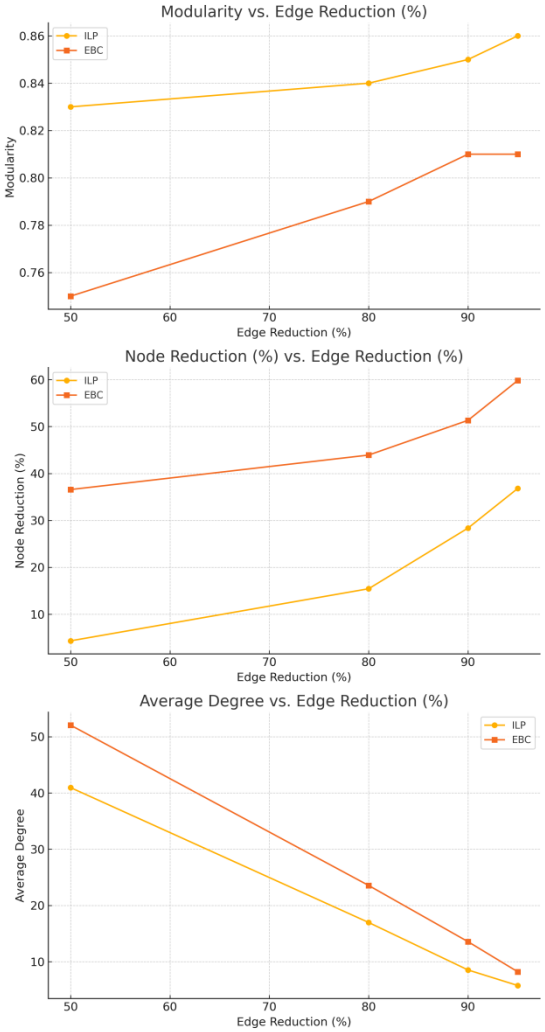

Figure S4. Comparative evaluation of ILP and EBC sparsification showing that ILP better preserves modularity, node count, and average degree across different edge reduction levels.

Section S5. Dataset Composition, Distribution, and Preprocessing

S5.1 Dataset Overview

The MOF dataset used in this study is extracted from the CSD MOF subset and includes a total of 14,296 experimentally validated MOFs. Each MOF is characterized by its organic linker, metal type, and key geometric descriptors, namely the Pore Limiting Diameter (PLD), Largest Cavity Diameter (LCD), and Largest Free Sphere (LFS). The dataset exhibits high structural diversity, featuring 53 unique metal types and 3,193 distinct linkers, covering a wide range of structural motifs relevant to MOF chemistry. The pore-related descriptors span from non-porous to highly porous structures, with PLD ranging from 0.0 Å to 71.5 Å, LCD from 0.006 Å to 71.64 Å, and LFS from -0.002 Å to 71.64 Å.

A summary of the dataset's key statistics and structural characteristics is provided in **Table S5.1**, which serves as the foundation for constructing MOFGalaxyNet**.**

Table S5.1. Summary of dataset composition and pore descriptor ranges.

Property	Value
Total MOFs	14,296
Number of unique metals	53
Number of unique linkers	3,193
PLD range	0.0 Å to 71.5 Å

LCD range	0.006 Å to 71.64 Å
LFS range	-0.002 Å to 71.64 Å

S5.2 Distribution Characteristics

The distribution analysis of key structural and compositional properties of the MOF dataset is summarized in **Figure S5**. As shown, the PLD, LCD, and LFS distributions exhibit heavy-tailed behavior, reflecting the coexistence of both microporous and mesoporous frameworks. The majority of MOFs possess pore-limiting diameters (PLD) below 10 Å, highlighting the dominance of small to medium pore structures. Furthermore, the distribution of metal types shows a skewed pattern, with transition metals such as Zn, Cu, Co, and Ni being the most frequently occurring, which is consistent with typical MOF synthesis trends. This distributional information confirms the structural diversity of the dataset, ensuring the generalizability of subsequent network analysis and machine learning tasks.

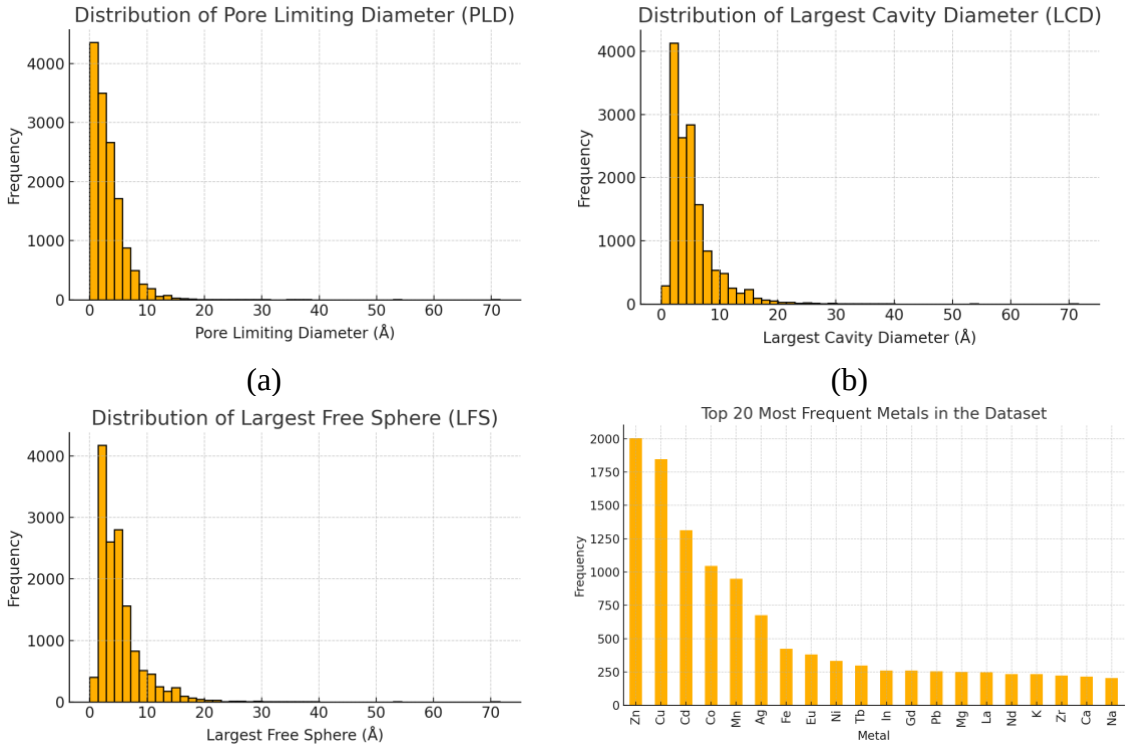

Figure S5: Distributions of pore sizes and metal occurrence highlighting the structural diversity of the MOF dataset.

S5.3 Preprocessing

To ensure the reliability and consistency of the dataset, we applied a systematic preprocessing procedure. First, MOFs with physically invalid pore descriptors (PLD, LCD, or LFS less than 0 or below 0.01 Å) were treated as outliers and removed. Second, MOFs with missing essential information, such as linker SMILES or metal descriptors, were excluded. Finally, all pore-related descriptors were normalized using Min-Max scaling to map them into the [0,1] range, making them suitable for network construction and machine learning tasks. This preprocessing step guarantees that the constructed MOFGalaxyNet is free from numerical artifacts and extreme values while retaining the structural diversity of the MOF dataset.

S5.4 Example of Raw Dataset

Table S5.2 presents example entries from the MOF dataset used to construct MOFGalaxyNet, showcasing the structure of the raw data. Each MOF is identified by its unique CSD reference code and includes its linker in SMILES format, the associated metal type, and key geometric descriptors: Largest Cavity Diameter (LCD), Pore Limiting Diameter (PLD), and Largest Free Sphere (LFS). These examples illustrate the typical information extracted from the CSD MOF subset and used during graph construction and analysis.

Table S5.2: Example MOF entries with linkers, metals, and pore descriptors from the CSD dataset.

refcode	linker SMILES	metal	LCD (Å)	PLD (Å)	LFS (Å)
ABAVIJ	OC(=O)c1ccncc1	Co	4.44	2.50	3.98
ABAVOP	OC(=O)c1ccncc1	Co	3.53	2.44	3.51
ABAVUV	OC(=O)c1ccncc1	Co	5.00	4.30	4.98

S5.5 Link to Previous Study

The MOF classes and pore size distributions are consistent with our previous work [1] (Jalali et al., *Journal of Cheminformatics*, 2023), confirming the generalizability of the selected subset.

References

(1) Jalali, M.; Wonanke, A. D.; Wöll, C. MOFGalaxyNet: a social network analysis for predicting guest accessibility in metal–organic frameworks utilizing graph convolutional networks. *Journal of Cheminformatics* **2023**, *15* (1), 94.