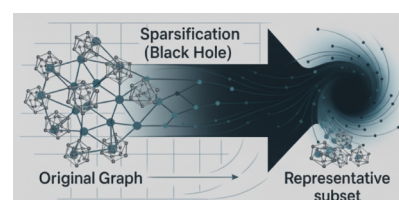


The Black Hole Strategy: Gravity-Based Representative Sampling for Frugal Graph Learning on Metal–Organic Framework Networks

Mehrdad Jalali,* A. D. Dinga Wonanke,* Pascal Friederich, and Christof Wöll

ABSTRACT: The expansion of large-scale materials databases has facilitated the development of graph-based representations, encoding structural and functional similarities as edges in data-driven networks. These enable machine learning models to leverage both local features and global relationships. However, densely connected datasets often introduce redundancy and noise, escalating computational complexity without improving performance. Here, we introduce the Black Hole Strategy, a gravity-based representative sampling method that constructs compact, informative subsets from large materials datasets while preserving essential structural and property diversity. Using metal–organic frameworks (MOFs) as a case study, we demonstrate that graph neural networks (GraphSAGE, GCN, and GAT) trained on Black Hole–sparsified datasets achieve comparable or superior classification and regression performance compared to full-dataset models, despite utilizing significantly fewer data points and reduced memory and training time requirements. Analysis of class-level confusion matrices confirms that critical structure–property relationships—such as pore-limiting diameter (PLD)—persist under substantial sparsification. An ablation study on gravity score weights validates the balanced formulation and robustness of the approach. Topological and efficiency benchmarks further demonstrate that the method preserves modularity, diversity, and connectivity across sparsification levels. These findings establish the Black Hole Strategy as a principled and frugal approach for machine learning in materials science, enabling efficient, interpretable, and scalable discovery workflows. Importantly, this work contributes to the objectives of the FAIRmat consortium, which aims to develop a FAIR data infrastructure for condensed matter physics and materials science. Our approach advances FAIR (Findable, Accessible, Interoperable, Reusable) data practices through optimized sampling techniques that enhance data management, reusability, and interoperability in materials informatics.



1. INTRODUCTION

The discovery and development of advanced materials have historically been major contributors to technological breakthroughs ranging from semiconductors for electronics to composite materials for aerospace applications. In recent years, the accelerating pace of technological innovation has intensified the demand for materials with unprecedented combinations of properties, including high strength, lightweight structures, selective reactivity, and tunable porosity. However, traditional experimental approaches are increasingly constrained by the immense combinatorial complexity arising from variations in chemical composition, structural design, processing conditions, and reaction environments.

Consequently, the integration of artificial intelligence (AI), machine learning (ML), and high-throughput computational techniques into materials science has emerged as a transformative strategy for guiding the discovery and optimization of next-generation materials.^{1,2} These data-driven approaches enable the rapid screening of vast chemical spaces, the identification of structure–property relationships, and the prediction of novel compounds with tailored functionalities,

which significantly reduces the time and cost associated with traditional trial-and-error methods in the field.

Metal–organic frameworks (MOFs) and high-entropy alloys (HEAs) represent two classes of materials characterized by extraordinary structural and compositional diversity. This complexity offers vast opportunities for applications but also creates significant challenges for rational design and discovery.^{3,4} To address the combinatorial explosion inherent in such systems, machine learning approaches must be combined with intelligent sampling strategies that can extract representative, high-quality subsets from large and heterogeneous chemical spaces. In this study, we focus on MOFs as a model system to demonstrate how this strategy can enhance AI-driven property prediction and facilitate more generalizable materials modeling.

MOFs are porous, crystalline materials composed of repeating units of well-defined metal and organic building blocks.⁵ Over the past two decades, MOFs have garnered increasing scientific interest for applications in gas storage, catalysis, drug delivery, sensing, energy storage, and gas separation.^{6–8} This surge in scientific interest has resulted in the synthesis and characterization of more than 120 000 unique MOFs, which have been deposited in the Cambridge Structural Database (CSD), along with millions of high-performance theoretically predicted hypothetical structures.^{9–11} Consequently, the chemical design space of MOFs has expanded to a scale that exceeds the capabilities of traditional computational and experimental methods.

To overcome this challenge, the integration of curated databases, high-throughput screening, and machine learning tools has emerged as a powerful alternative for accelerating MOF discovery and property prediction.^{12–16} High-throughput computational workflows combined with machine learning (ML) models have enabled the rapid prediction of key material properties and facilitated the identification of promising candidates for targeted applications. Descriptor-based approaches have further accelerated discovery pipelines by enabling the quantitative correlation of structural features with key performance metrics such as gas adsorption capacity, catalytic activity, and electronic properties.^{17–19} Moreover, the rise of materials genomics has empowered researchers to leverage big data analytics for comprehensive structure–property evaluations.^{1,3} Recent advances in large language models (LLMs) fine-tuned for chemical reasoning have further democratized access to MOF research by enabling broader engagement even in data-scarce settings.²⁰

Despite substantial progress in machine learning for materials science, the accumulation of redundant and noisy data continues to slow the development of efficient models. As datasets grow larger, they increasingly accumulate structurally peripheral and low-information data points, which obscure meaningful structure–property relationships and prevent model generalization. This informational redundancy not only increases computational costs but also slows the pace of materials discovery and optimization. To overcome this bottleneck, it is necessary to develop not only sophisticated machine-learning architectures but also principled data sampling strategies that can extract representative and information-rich subsets while systematically eliminating redundant, noisy, and structurally peripheral data points.

Graph-based learning frameworks offer a promising pathway for managing the intrinsic complexity of large-scale MOF datasets.²¹ In such frameworks, crystal structures are represented as graphs with nodes corresponding to atoms or structural units, and edges representing chemical bonds or topological relationships.²² This representation enables machine learning models to capture both local atomic environments and global structural connectivity. Recent applications of graph neural networks (GNNs) have demonstrated strong progress in predicting properties such as CO₂ capture performance, electronic structure, and partial atomic charges. Yet, most studies still treat MOFs as isolated graphs, overlooking relational similarities across the broader design space. Notably, Chen et al. developed a graph-based deep learning increment theory for enthalpy prediction with strong generalization to larger molecules,²³ while Muthiah et al. applied GNNs to chemical property prediction in process systems—further underscoring the versatility of graph-based learning across chemical domains.

Traditional machine learning approaches often treat each material as an independent data point. In contrast, we propose viewing the entire dataset as a graph, where each material is represented as a node and edges capture structural or chemical similarities. We refer to this abstraction as a material network.^{14,15} This network-based perspective that has been successfully applied in our previous work on high-entropy alloys (HEAs)^{25,26} is extended here to MOFs. It enables the use of graph theory tools—such as community detection and centrality measures—to better capture redundancy, diversity, and structural coverage across large materials datasets.

To address these challenges, we previously introduced the concept of social network analysis as a framework for investigating the interrelationships among MOFs, enabling the design of predictive models grounded in structural interactions across the chemical space. In this previous study, we represented the entire MOF design space as a single cohesive graph, where each node corresponded to an individual MOF and edges quantified the structural similarity between them. Through this networked representation, we discovered that the multilayered interactions inherent in social networks could be strategically harnessed to eliminate redundancies and yield a more efficient and structurally diverse representation.

Building upon these findings, we introduce the Black Hole Strategy as a generalized and scalable sampling framework for extracting representative metal–organic frameworks (MOFs) from large graph-based datasets.

For this reason, we decided to develop a scalable and generalizable approach for constructing lightweight, information-rich datasets that can support efficient and accurate machine learning. Hence, we introduce the Black Hole Strategy, which is a principled sampling framework designed to systematically identify and retain representative nodes within large material networks while minimizing redundancy. As proof of principle, we apply this framework to a MOF social network and evaluate its effectiveness by training and testing graph-based machine learning models (GraphSAGE, GAT, and GCN) using pore-limiting diameter (PLD) classification as a representative task. Our findings show that strategic network sparsification can maintain and even enhance predictive performance.

2. THEORY AND METHODOLOGY

2.1. The Black Hole Strategy: Conceptual Foundations and Definitions. The Black Hole (BH) Strategy, introduced in this work, is a gravity-based representative sampling approach designed to identify and retain the most influential nodes and edges in a materials graph. The name reflects its conceptual analogy to celestial black holes—regions in space characterized by an intense gravitational pull that attracts and concentrates surrounding matter. Similarly, in our method, nodes with higher graph influence—quantified through a gravity-inspired metric that combines structural centrality, property similarity, and connectivity—exert a stronger “pull” on their neighboring nodes and edges. By prioritizing these influential nodes, the method constructs a compact yet structurally and functionally representative subset of the original dataset, reducing redundancy and computational overhead while maintaining essential structure–property relationships.

In a given graph $G = (V, E)$, a node $v \in V$ is classified as a Black Hole Node if its gravity score $g(v)$ exceeds an empirically defined threshold γ , i.e., $g(v) \geq \gamma$.

These nodes are highly connected and structurally diverse, forming a distilled subset of the original dataset that captures the

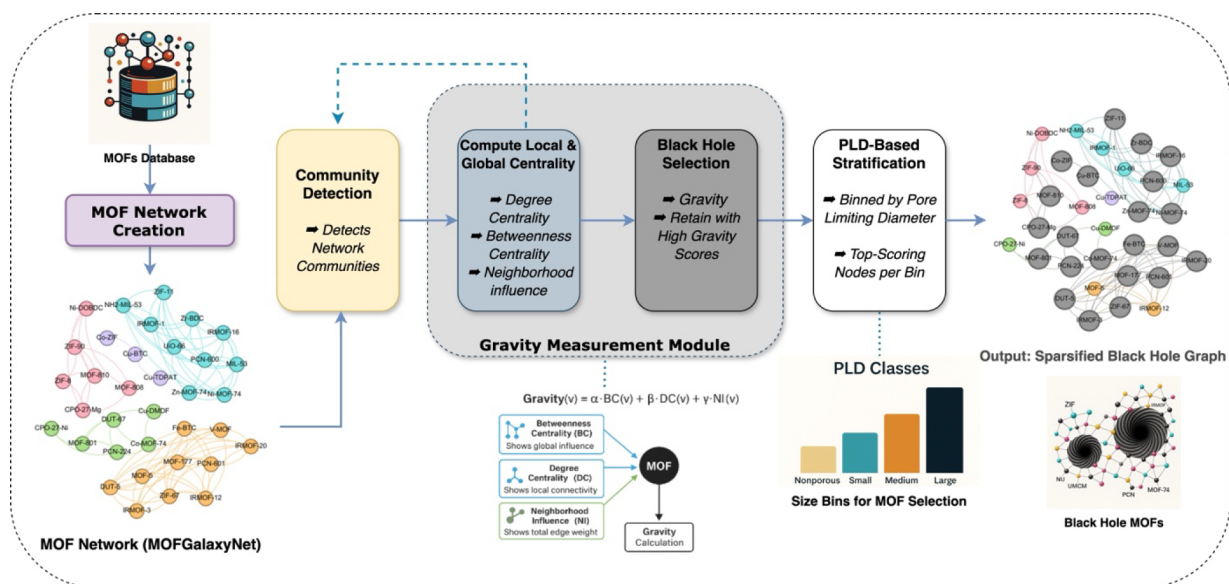


Figure 1. Workflow of the Black Hole Strategy. MOF networks are constructed and clustered into communities. Nodes are scored by composite gravity metrics derived from centrality measures. High-gravity nodes are selected and stratified by PLD classes to form a sparsified, property-aware MOF graph for efficient machine learning.

essential functional and geometric variation needed for machine learning applications.

Definition 1 (Black Hole Strategy): The Black Hole Strategy is a centrality-driven, property-aware sampling framework designed to extract a compact, representative subset of nodes from large-scale graphs. It systematically eliminates low-gravity nodes and weakly weighted edges while preserving core community structure and property diversity.

Definition 2 (Gravity Score): The gravity score is a quantitative measure of a node’s overall influence or importance within a network. It integrates multiple structural metrics to reflect both the local and global relevance of a node. Higher gravity scores indicate nodes that are more central, better connected, and more structurally significant in the network topology.

Definition 3 (Sparsification Threshold τ): The pruning rate $\tau \in (0,1)$ is a user-defined parameter that controls the degree of network reduction. It determines the fraction of nodes with the lowest gravity scores (within each community) and the weakest edges (globally or locally) to be removed during the sparsification process.

A summary of the Black Hole Strategy is illustrated in [Figure 1](#).

The first step in our Black Hole Strategy involves constructing a social network, following the approach introduced in our previous work.²² To support the structural foundation of our analysis, we utilized MOFGalaxyNet, a large-scale similarity-based network of 14 000 MOFs constructed from curated structural and chemical descriptors. Each node represents a MOF, and edges encode pairwise similarity relationships derived from a dedicated scoring engine. The network’s high modularity enables precise community detection and pattern extraction. The complete workflow for generating MOFGalaxyNet is described in detail in [Section S2](#).

To extract meaningful structural insights from MOFGalaxyNet, we apply social network analysis (SNA) in two key stages. First, community detection is performed to identify modular clusters of related MOFs, preserving the inherent structure of

the network while preventing bias toward densely connected regions. These communities serve as strata, enabling balanced and systematic sampling across the network. Second, we compute local and global centrality metrics—including degree centrality, betweenness centrality, and neighborhood influence (measured by the sum of edge weights)—to assess node-level importance. These metrics are normalized using min-max scaling and combined into a composite gravity score, identifying the most structurally influential nodes, analogous to gravitational centers in celestial systems.

This analysis feeds into the Gravity Measurement Module, where MOFs are ranked based on their gravity scores. In the Black Hole Selection stage, nodes with high gravity values are retained as structurally dominant candidates. To maintain material property diversity, a PLD-Based Stratification step is applied: MOFs are grouped into bins by pore-limiting diameter (PLD) (Nonporous, Small, Medium, Large), and the top-scoring MOFs from each bin are selected. This ensures that the final representative subset reflects both structural importance and diversity in material properties.

The Black Hole Strategy thus integrates network topology, centrality-based gravity scoring, and stratified sampling into a unified framework. The result is a sparsified black hole graph—a reduced yet representative subset of the original MOF network that retains modularity, connectivity, and diversity.

To improve robustness, the Black Hole Strategy incorporates a bootstrap step, where $|E|$ edges are sampled with replacement from the original graph before sparsification. This step is not strictly required for scalability (as MOFGalaxyNet is computationally manageable), but it enhances stability by accounting for stochasticity in community detection and ensuring reproducibility across multiple runs.

A pseudocode for the implementation of the Black Hole Strategy is summarized in [Algorithm 1](#). This workflow ensures the selection of an optimal subset of MOFs while preserving network diversity and structural integrity. This algorithm can be implemented on a social network to identify the most representative and unique nodes within the network.

```

Input:
E: Weighted edge list of the MOF similarity network
S: Node summary table, including structural descriptors and PLD values
 $\tau \in (0, 1)$ : Sparsification threshold (fraction of nodes and edges to prune)
R: Number of bootstrap runs
 $\alpha, \beta, \gamma$ : Weights for centrality metrics (Degree, Betweenness, Neighborhood Influence)

Output:
E_BH: List of sparsified edge sets  $\{E_{BH}^1, \dots, E_{BH}^R\}$ 
NodeFeatures: Node-level table with normalized centralities, gravity scores, and PLD class labels

for each run  $r$  in  $\{1, \dots, R\}$  do
    1. Bootstrap sampling:
        - Sample  $|E|$  edges with replacement from  $E$  to construct graph  $G = (V, E_r)$ 

    2. Community detection:
        - Detect communities  $C = \{C_1, \dots, C_k\}$  using the Louvain algorithm

    3. Centrality computation for each community  $C_i$ :
        for each node  $v$  in  $C_i$  do
            - Compute Degree Centrality:  $DC(v)$ 
            - Compute Betweenness Centrality:  $BC(v)$ 
            - Compute Neighborhood Influence:  $NI(v) = \sum w(u,v) \ \forall (u,v) \in E_i$ 

        - Normalize DC, BC, and NI within  $C_i$  using Min-Max scaling:
             $DC_{norm}(v), BC_{norm}(v), NI_{norm}(v)$ 

    4. Gravity score calculation:
        for each node  $v \in V$  do
            -  $Gravity(v) = \alpha \cdot DC_{norm}(v) + \beta \cdot BC_{norm}(v) + \gamma \cdot NI_{norm}(v)$ 

    5. Property-aware binning:
        - Assign each node to a PLD bin:
            {Nonporous, Small, Medium, Large}

    6. Pruning within PLD bins:
        for each bin  $b$  do
            - Sort nodes by  $Gravity(v)$  in descending order
            - Select top  $(1 - \tau)\%$  of nodes in bin  $b$ 

    7. Subgraph reconstruction:
        - Define  $V_{BH}$  = selected nodes from all bins
        - Define  $E_{BH}^r$  = edges from  $E_r$  connecting nodes in  $V_{BH}$ 
        - Construct subgraph  $G_{BH} = (V_{BH}, E_{BH}^r)$ 

    8. Save results:
        - Append  $E_{BH}^r$  to  $E_{BH}$ 
        - Store NodeFeatures with centrality scores,  $Gravity(v)$ , and PLD bin

end for

```

Algorithm 1 Steps of the Black Hole Strategy, detailing the creation of social network, community detection, gravity scoring, and stratified sampling for optimal MOF selection.

2.2. Community Detection. Community detection in the MOF network was performed using the Louvain algorithm,²² which is widely adopted for large-scale networks due to its scalability and ability to identify high-modularity partitions. The Louvain algorithm operates in two iterative phases: (i) local movement of nodes, where each node is reassigned to the community of one of its neighbors if such a move increases modularity, and (ii) community aggregation, where identified communities are collapsed into supernodes to construct a reduced network. These two steps are repeated until modularity converges, resulting in a hierarchical community structure. This approach provides an efficient alternative to edge-removal-based

methods (e.g., Girvan–Newman) and is particularly suited for sparsified graphs, where preserving modular organization is crucial for representative sampling. For a detailed explanation of the community detection process, including the mathematical formulations and iteration specifics, refer to [Section S3](#).

2.3. Gravity Scoring. The significance of each node in the network is quantified by its gravity score, computed as a weighted sum of three normalized centrality measures: Degree Centrality (DC), Betweenness Centrality (BC), and Neighborhood Influence (NI), the latter representing the total weight of edges connected to the node. The mathematical formulation is given by the following:

$$\text{Gravity}(v) = \alpha \cdot \text{DC}_{\text{norm}}(v) + \beta \cdot \text{BC}_{\text{norm}}(v) + \gamma \cdot \text{NI}_{\text{norm}}(v) \quad (1)$$

Here, α , β , and γ are the weights assigned to Degree Centrality (DC), Betweenness Centrality (BC),²⁴ and Neighborhood Influence (NI), respectively. In this study, we assign equal weights to all three components. To examine the relative contribution of these terms, we conducted an ablation study varying α , β , and γ while maintaining $\alpha + \beta + \gamma = 1$. As summarized in Table S1 (Section S4), emphasizing a single component (e.g., $\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$) yielded marginal improvements for specific models, but results fluctuated across sparsification levels. By contrast, the balanced configuration ($\alpha = \beta = \gamma = 0.33$) achieved consistently stable and competitive accuracy across GAT, GCN, and GraphSAGE. This validates our choice of equal weighting in all main experiments, as it avoids bias toward any single factor, preserves interpretability, and ensures robustness across diverse GNN architectures.

Overall, this formulation ensures a balanced contribution from each metric while highlighting the role of neighborhood influence, which captures the interaction strength of nodes within their communities—a critical property in weighted networks. By combining these three structural features, the gravity score provides a comprehensive measure of node importance, serving as the foundation of the Black Hole Strategy. For detailed numerical results and extended discussion, see Section S4.

2.4. Preserving Material Property Diversity through Stratified Sampling. In this study, we used the pore-limiting diameter (PLD) as a key material property to validate the effectiveness of our Black Hole data sparsification approach. Since the PLD values are continuous, we converted them into categorical classes, as shown in Table 1. This enabled a more

Table 1. Transformation of Pore-Limiting Diameter (PLD) Values from Continuous to Categorical Variables

Classes	PLD Range (Å)
Nonporous	PLD $\leq$ 2.4
Small	2.4 < PLD $\leq$ 4.4
Medium	4.4 < PLD $\leq$ 5.9
Large	PLD > 5.9

structured approach to incorporating stratified sampling, ensuring that the reduced network maintains a balanced representation of PLD categories while prioritizing the most structurally significant nodes.

To illustrate the effectiveness of the Black Hole Strategy, we applied it to a sample network of 35 metal–organic frameworks (MOFs) segmented into five communities. As shown in Figure 2, the method systematically reduces the network by removing

structurally less influential nodes while retaining key representatives—referred to as Black Hole Nodes—based on gravity scores. This progressive pruning preserves modularity and connectivity, emphasizing structurally significant MOFs across all communities.

A detailed explanation and full-resolution visualization of this reduction process are provided in Section S5.

3. RESULTS AND DISCUSSION

Applying the Louvain community detection algorithm to the MOF social network, which contains 14 000 MOFs, yielded 836 distinct communities with a total modularity score of 0.834. This modularity score measures the robustness and effectiveness of the Louvain algorithm in accurately identifying distinct communities within the social network. This indicates that communities detected are significantly more interconnected internally than would be expected by random chance. The MOF social network and the identified communities are illustrated in Figure 3. Here, only highly populated communities are illustrated for better visualization, with colors representing different communities.

To further characterize the 836 identified communities, we employed a combination of gravity-based metrics, network centrality measures, and stratified sampling analysis to select representative metal–organic frameworks (MOFs) from each community. This approach ensured that the nodes selected as centroids from each community accurately encode the structural and functional diversity inherent to each community. Furthermore, the number of centroids extracted from each community depends on the percentage of sparsification applied during centroid analysis. The variation in the number of nodes representing the exact number of MOFs with varying percentages of sparsification is illustrated in Figure 4.

Figure 4 illustrates that once the sparsification threshold increases, the number of retained MOFs decreases steadily. This consistent reduction reflects the direct effect of the Black Hole Strategy, which prunes structurally less informative nodes to yield a compact, representative subset. The resulting compression significantly reduces network size while preserving diversity through controlled, gravity-based selection. Notably, although the original MOFGalaxyNet contains approximately 14 000 MOFs, the node count at the sparsification threshold 0 begins at around 12 000. This discrepancy arises because some MOFs were excluded during the construction of the initial graph, specifically those lacking sufficient structural similarity (i.e., edge weights below the defined threshold) and thus became isolated nodes. These unconnected structures were removed to ensure a cohesive and meaningful network representation.

To evaluate the effectiveness and robustness of this data reduction, we further assessed whether key structural and



Figure 2. Progressive sparsification of a sample MOF network from left (dense) to right (sparse), highlighting retained high-gravity nodes as representative MOFs.

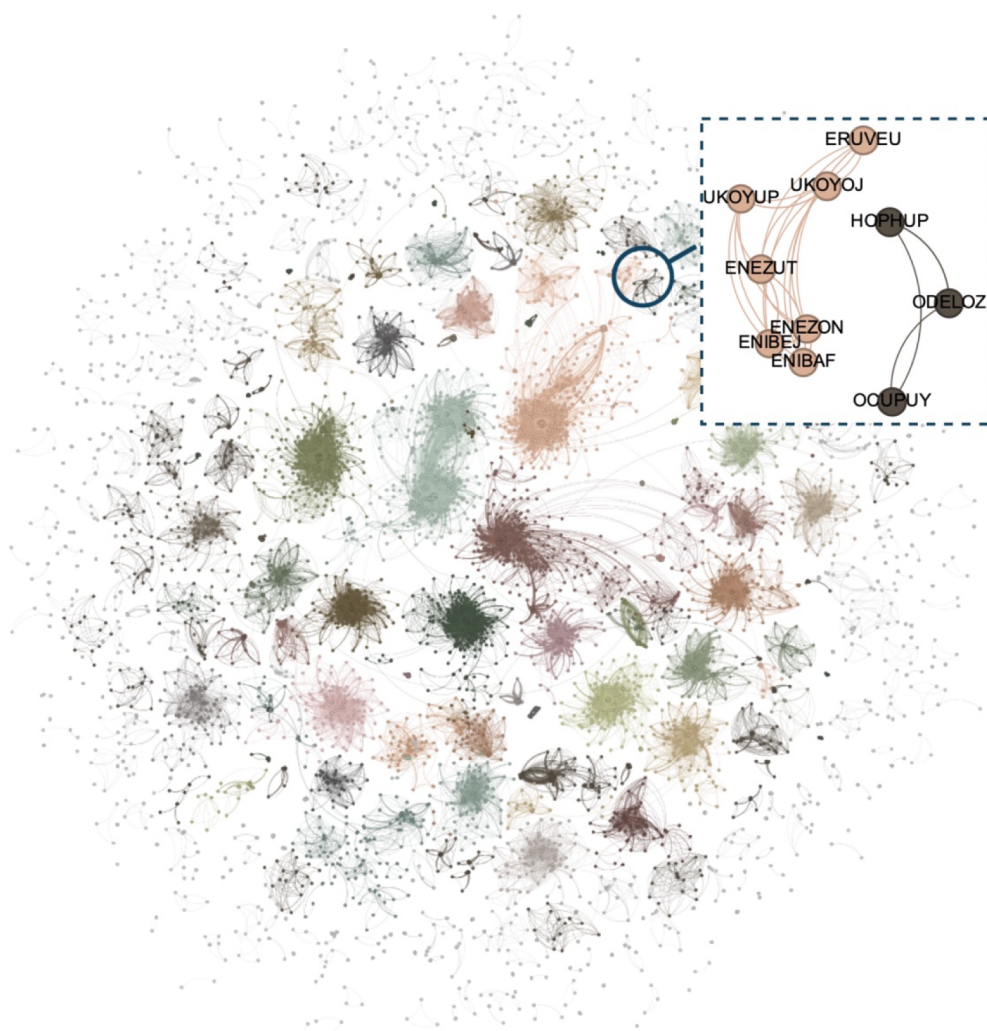


Figure 3. MOF social network visualized using Gephi,²³ where each dark node represents a MOF and colors represent distinct communities. Smaller communities are omitted for clarity. Inset: a zoomed view of a selected community with internal connections.

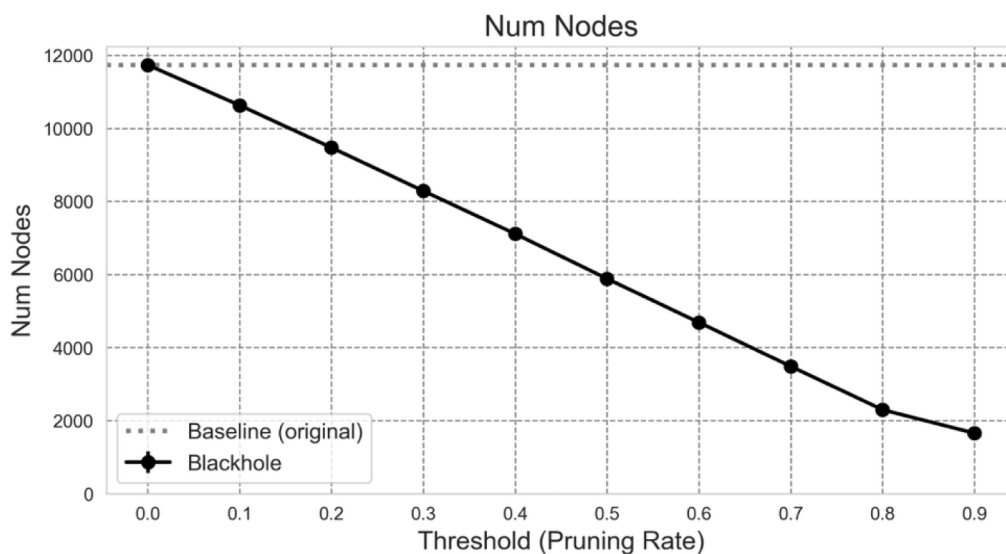


Figure 4. Number of representative MOFs extracted from the social network at different sparsification levels.

geometric properties, such as pore-limiting diameter (PLD), metal distribution, and linker types, were preserved across different sparsification levels. These comparisons were per-

formed within communities to ensure that the selection of representative MOFs remained consistent and structurally

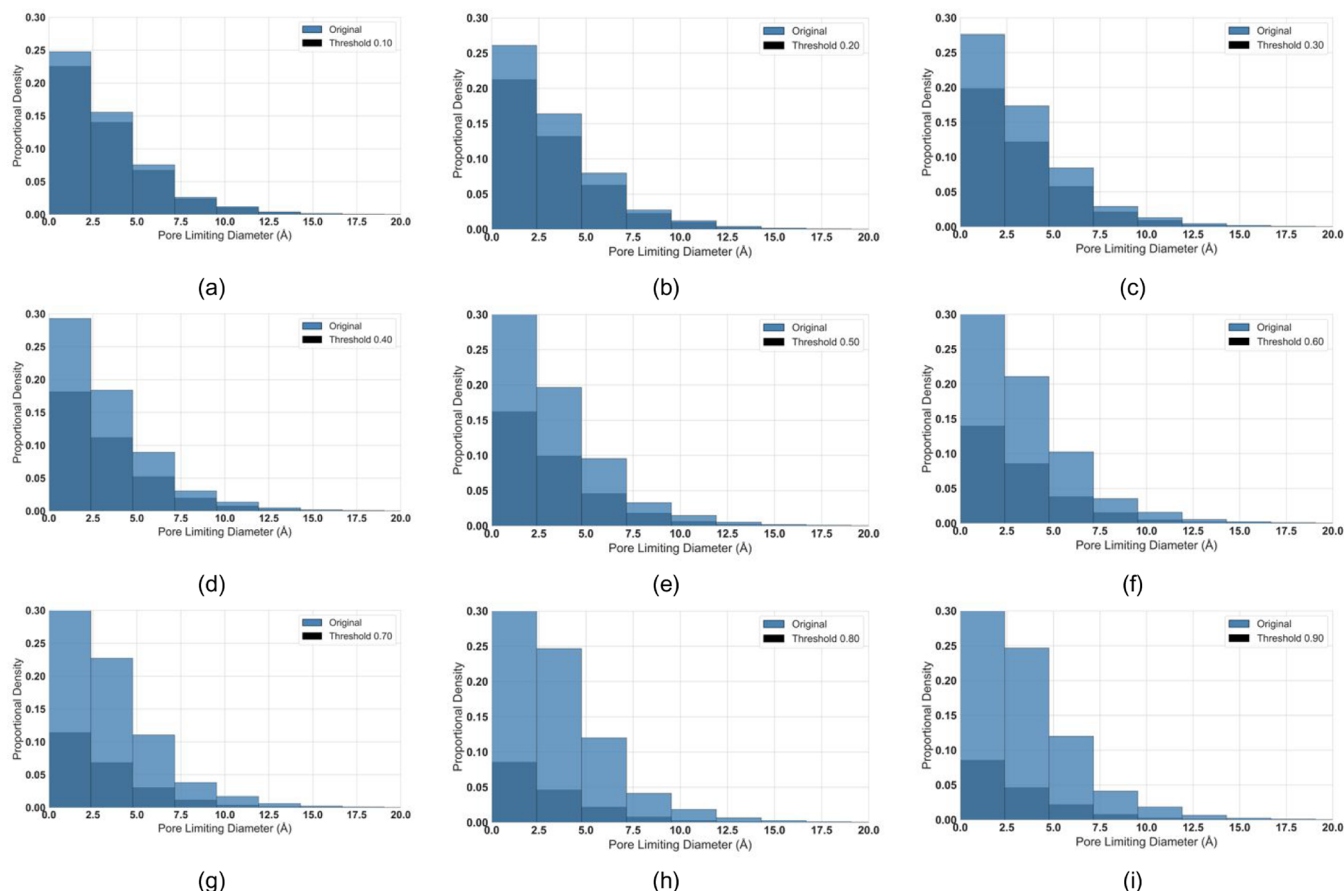


Figure 5. PLD distribution comparison between the original MOF dataset (light blue) and Black Hole–selected subsets (dark blue) across sparsification thresholds. Panels: (a) 10%, (b) 20%, (c) 30%, (d) 40%, (e) 50%, (f) 60%, (g) 70%, (h) 80%, (i) 90%.

meaningful, even under varying degrees of edge and node removal.

3.1. Comparison of PLD Representativeness across Communities at Varying Sparsification Levels. Figure 5 presents the distribution of pore-limiting diameter (PLD) values for MOFs in the original dataset alongside those selected as community representatives under various sparsification levels. Across all sparsification stages, the representative subsets exhibit PLD distributions that closely mirror the original, indicating that the sparsification process preserves the geometric diversity of the dataset. This alignment demonstrates the robustness of the Black Hole selection strategy in maintaining structural representativeness even as the graph is pruned. Notably, up to 70% sparsification, the representative distributions remain highly consistent, with only minimal deviations. At 90% sparsification, while the core characteristics are still captured, a modest shift is observed in the upper PLD range (above ~ 7 Å), suggesting that extreme sparsification may begin to affect the fidelity of structural representation in some communities.

3.2. Comparison of Metal Atoms of Top Five Representativeness across Communities at Varying Sparsification Levels. Figure 6 presents the distributions of the top five most frequently occurring metal atoms in MOFs from the original dataset, compared with those of the representative MOFs selected from each community at varying levels of sparsification. At lower sparsification levels (up to 20%), the metal distributions of the representative sets remain closely aligned with those of the original dataset, indicating strong representational fidelity. However, as sparsification increases,

noticeable deviations begin to emerge. These deviations suggest a growing sensitivity in preserving less common metal species, likely due to the removal of peripheral or weaker connections that are essential for retaining structural diversity. At higher sparsification levels (e.g., 70–90%), the representative selection increasingly favors the most dominant metals, resulting in a narrower compositional profile. This trend highlights a trade-off between graph sparsity and chemical representativeness, emphasizing the importance of balancing sparsification with diversity preservation in MOF selection strategies.

3.3. Comparison of Organic Linkers of Top Five Representativeness across Communities at Varying Sparsification Levels. Figure 7 illustrates the distributions of the top five most frequently occurring organic linkers—converted from SMILES to IUPAC names—in the original dataset and in the representative MOFs selected from each community at varying sparsification levels. Compared to metal atoms and PLD distributions, organic linkers exhibit markedly higher sensitivity to sparsification. At the lowest sparsification level (10%), the distribution is relatively well preserved; however, even modest increases in sparsity result in growing deviations from the original linker profile. This trend reflects the greater structural diversity and network fragility associated with organic linkers: as sparsification removes more edges, MOFs containing less common linkers are more likely to be excluded. Consequently, the selection process begins to favor dominant linker motifs, leading to a compressed and less chemically diverse representation. These findings suggest that while sparsified networks can retain geometric and topological

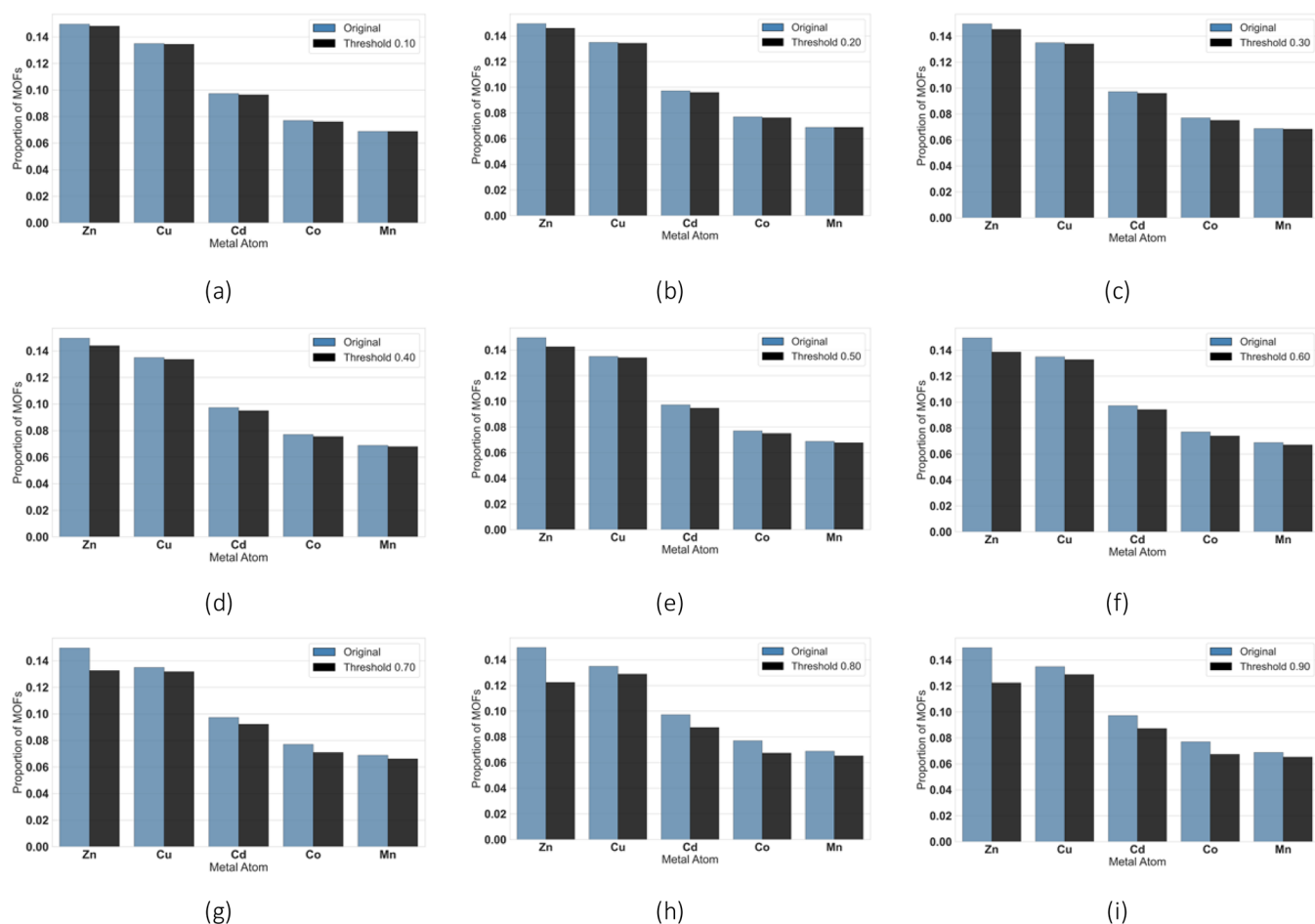


Figure 6. Distribution of the top five metal atoms in the original MOF dataset (blue) and in Black Hole–selected representative subsets (black) across sparsification levels from 10% to 90% (a–i). Lower sparsification levels preserve metal diversity well, while higher levels increasingly favor dominant metals.

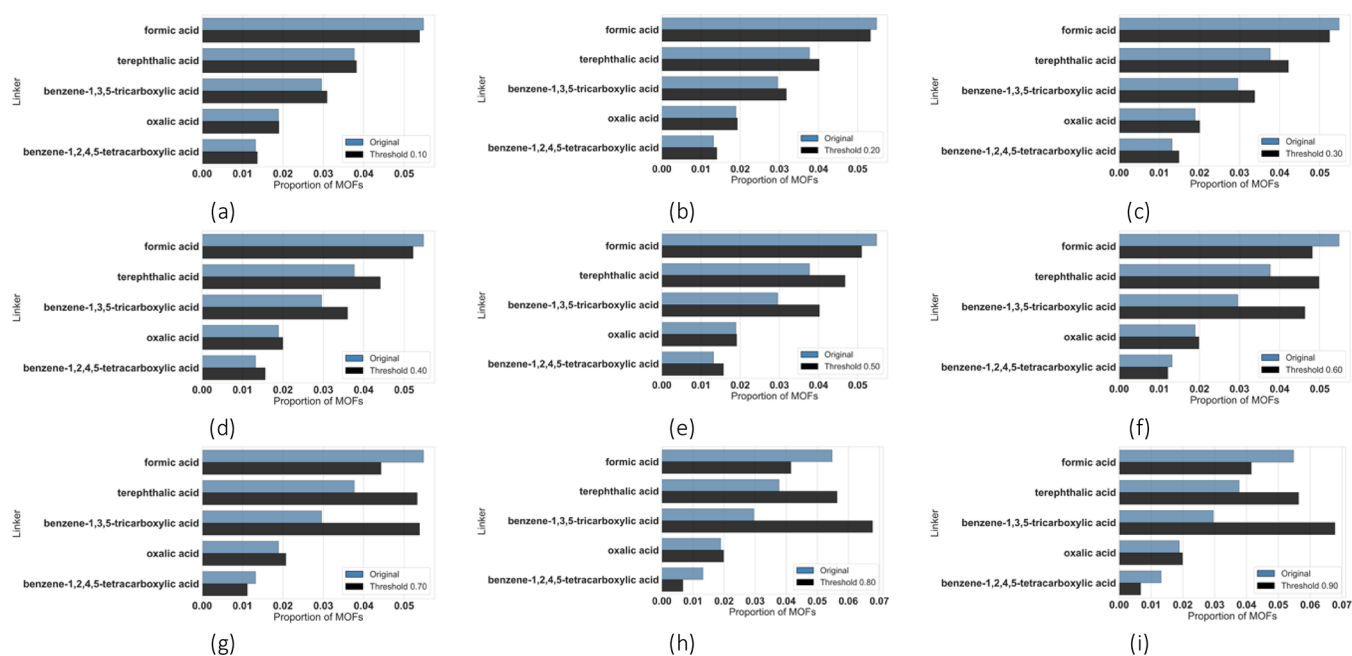


Figure 7. Distribution of the top five organic linkers in the original MOF dataset (blue) and in Black Hole–selected subsets (black) across sparsification levels from 10% to 90% (a–i). While linker diversity is mostly preserved at low sparsification, deviations increase at higher thresholds, with dominant linkers becoming more prevalent.

features, preserving the chemical diversity, particularly among organic linkers, requires maintaining higher connectivity in the graph.

Together, the analyses of pore-limiting diameter (PLD), metal atom composition, and organic linker distribution demonstrate that the Black Hole Strategy offers a robust and efficient framework for selecting representative MOFs from large-scale datasets. At sparsification levels below 30%, the representatives selected from each of the 836 communities effectively preserve the intrinsic geometric features of the full dataset, with minimal loss in PLD distribution. Metal composition is also well maintained at these levels, while organic linker diversity—being more structurally sensitive—requires higher connectivity to remain faithfully represented. Overall, the Black Hole Strategy proves to be a powerful sampling method for extracting a compact yet chemically and geometrically diverse subset of MOFs. This makes it especially well-suited for high-throughput screening and large-scale materials discovery, where balancing representational fidelity with computational efficiency is critical.

3.4. Evaluation of the Black Hole Strategy with Graph Neural Networks. To evaluate the Black Hole Strategy, we utilized graph neural networks (GNNs), which are well-suited for analyzing networks where meaningful relationships exist between nodes. In our MOF graph, each node represents a complete metal–organic framework (MOF), and edges indicate structural similarity between MOFs. Traditional machine learning treats data points separately and often misses these relationships as an inter-MOF structure. GNNs, however, can learn from both the individual properties of each MOF and the similarities between them by having both local features and global topology. This makes GNNs ideal for evaluating how effectively the Black Hole Strategy retains important information while reducing the network size.

In this study, classification was conducted based on pore-limiting diameter (PLD), a key structural descriptor of MOFs. The continuous PLD values were categorized into four discrete classes—Nonporous, Small, Medium, and Large—as defined in Table 1. These categories were used as target labels for all GNN-based prediction tasks. This choice ensures that the evaluation of the Black Hole Strategy directly reflects its ability to preserve structurally relevant information required for accurate classification of PLD-based properties.

Three types of GNN models were employed to test the method. Graph Attention Networks (GAT) learn which neighboring MOFs are most important by using attention weights. Graph Convolutional Networks (GCN) combine information evenly from each MOF’s neighbors. GraphSAGE utilizes sampling to efficiently work with large graphs, learning by aggregating features from nearby MOFs. These models enable evaluation of the sampling strategy’s effectiveness under various learning styles and network sizes. The specifics of these algorithms are provided in Section S6.

The Black Hole Strategy was evaluated against three baseline sparsification methods: stratified sampling,²⁷ PageRank-based selection,²⁸ and K-Center (farthest-point) sampling.²⁹ Stratified sparsification preserves dataset diversity by sampling MOFs from distinct property groups, such as pore-limiting diameter (PLD) categories. To strengthen this baseline, we implemented a customized variant that (i) guarantees the inclusion of key MOFs used in testing and (ii) leverages the structure of the MOF graph to guide edge selection, ensuring a more balanced and informative comparison.

In addition, we incorporated PageRank sparsification, a centrality-driven strategy that prioritizes nodes with higher global importance in the network, and K-Center sparsification, which represents a clustering/coverage approach by selecting nodes that maximize coverage of the graph through farthest-point representatives. These methods complement stratified sampling by reflecting different design principles—diversity preservation, centrality-based influence, and coverage maximization—thereby providing a stronger and more comprehensive evaluation framework.

All methods, including Black Hole, were executed 10 times per sparsification level to assess consistency and account for stochastic variation. Community detection for stratified sampling was performed using the Louvain algorithm, whose nondeterministic nature naturally introduces variation across runs, enabling stability analysis. To ensure fairness across methods and thresholds, a fixed representative test set was extracted from the original MOFGalaxyNet prior to sparsification and applied uniformly to all experiments. This ensured that identical test nodes were retained regardless of the sampling method or sparsification level. For stratified sampling, community detection was conducted with a resolution parameter of 0.5, favoring broader community structures well-suited to this task.

To guarantee test set availability at all sparsification levels, test nodes were selected from one of the most sparsified graphs, ensuring their presence in less sparsified versions as well. A limitation of this design is that, in denser graphs, some test nodes may become structurally isolated, with few or no edges. Since graph neural networks rely on neighborhood information, such isolation can reduce predictive accuracy. Nonetheless, this approach preserves consistency and comparability across all configurations.

For each sparsification threshold (ranging from 0.0 to 0.9), a fixed run index (Run = 0) was used to generate sparsified graphs, thereby removing variability from repeated sampling. Evaluation metrics were computed on the largest connected component of each graph to ensure structural comparability. This controlled setup guarantees that observed performance differences are attributable to the intrinsic properties of the Black Hole, stratified, PageRank, and K-Center strategies, rather than inconsistencies in the test set or graph topology.

All experiments were repeated for $R = 10$ bootstrap runs per sparsification level to assess stability. Performance variations between runs were negligible (typically <1% accuracy difference), and results converged after ~ 5 –7 runs. The averaged results reported in Table 2 and Figure 8 are therefore representative and robust, confirming that the Black Hole Strategy is not sensitive to the number of bootstrap replicates.

Table 2 presents the mean classification accuracies of three GNN models—GAT, GCN, and GraphSAGE—evaluated across pruning thresholds under four sparsification strategies: Black Hole, Stratified, PageRank, and K-Center. Each value represents the average over 10 independent runs, ensuring statistical robustness.

Across nearly all pruning levels, the Black Hole Strategy consistently outperforms the baseline methods. The performance gap is particularly notable at moderate sparsity levels (20–30%), where Black Hole retains structural representativeness more effectively, resulting in superior classification accuracy. For example, at 30% pruning, GCN with Black Hole achieves an accuracy of 0.783, compared to 0.732 with Stratified, 0.725 with PageRank, and 0.723 with K-Center.

Table 2. Mean Classification Accuracies of GAT, GCN, and Graphsage across Pruning Thresholds under Four Sparsification Strategies^a

Threshold	Black Hole	Stratified	PageRank	K-Center
(a) GAT				
0.0	0.781	0.733	0.727	0.720
0.1	0.791	0.730	0.720	0.721
0.2	0.809	0.727	0.729	0.717
0.3	0.805	0.730	0.719	0.721
0.4	0.776	0.706	0.702	0.721
0.5	0.768	0.689	0.695	0.718
0.6	0.765	0.681	0.688	0.710
0.7	0.770	0.683	0.690	0.707
0.8	0.771	0.681	0.689	0.705
0.9	0.775	0.679	0.688	0.703
(b) GCN				
0.0	0.833	0.723	0.746	0.755
0.1	0.801	0.733	0.740	0.743
0.2	0.802	0.736	0.743	0.751
0.3	0.783	0.732	0.725	0.723
0.4	0.774	0.710	0.706	0.726
0.5	0.760	0.702	0.698	0.718
0.6	0.756	0.699	0.695	0.716
0.7	0.754	0.696	0.692	0.714
0.8	0.752	0.695	0.691	0.713
0.9	0.751	0.694	0.690	0.712
(c) GraphSAGE				
0.0	0.780	0.731	0.731	0.731
0.1	0.781	0.726	0.726	0.726
0.2	0.776	0.722	0.722	0.722
0.3	0.798	0.727	0.727	0.727
0.4	0.789	0.709	0.709	0.709
0.5	0.780	0.702	0.702	0.702
0.6	0.774	0.698	0.698	0.698
0.7	0.771	0.696	0.696	0.696
0.8	0.770	0.694	0.694	0.694
0.9	0.768	0.693	0.693	0.693

^aEach value represents the mean of 10 independent runs. Bold highlights the best method at each threshold.

Among the models, GraphSAGE demonstrates the greatest robustness, maintaining accuracies near or above 0.74 even at 90% pruning under Black Hole, in contrast to a sharper decline with Stratified or PageRank. GAT and GCN exhibit steeper accuracy drops at mid-level sparsification when using stratified sampling, likely due to the removal of structurally important inter-MOF connections that Black Hole preserves.

Interestingly, performance across all models shows a dip around the 40–60% sparsification range, suggesting a transition zone where excessive edge removal fragments the graph and partially isolates test nodes. However, this degradation is consistently less severe under Black Hole than under the baselines. At higher sparsification (70–90%), accuracy improves again—particularly with Black Hole—likely because the remaining subgraphs form tighter, more informative clusters despite their reduced size.

In addition to the tabulated averages, Figure 8a–c visualizes the classification accuracies for each GNN model—GraphSAGE, GAT, and GCN—across all sparsification levels, including error bars representing the standard deviation from 10 runs.

Figure 8a shows that GAT models trained with the Black Hole Strategy consistently outperform all baselines, particularly

beyond 40% sparsification. While accuracy dips slightly around 50%, Black Hole maintains values above 0.7 even at 90% sparsification. In contrast, Stratified and PageRank methods suffer sharp declines below 0.5, indicating that random or centrality-biased removal disrupts the attention mechanism’s neighborhood focus.

Figure 8b highlights that GCN models are more sensitive to sparsity, but the Black Hole Strategy clearly mitigates this effect. At 70–90% sparsification, Stratified and PageRank collapse below 0.55 accuracy, whereas Black Hole sustains values around 0.7 with noticeably smaller variance. This stability emphasizes the importance of preserving structurally informative nodes in aggregation-based models.

Figure 8c demonstrates that GraphSAGE exhibits the strongest robustness overall. Even under extreme sparsification (90%), the Black Hole Strategy maintains an accuracy near 0.75, substantially higher than all other baselines. PageRank and Stratified sampling show pronounced degradation past 60%, while K-Center exhibits a gradual but consistent decline.

Taken together, these plots confirm that the Black Hole Strategy not only secures higher mean accuracies but also reduces run-to-run variability, offering stronger reproducibility. Its performance advantage is particularly evident in mid-to-high sparsification ranges, where conventional baselines fail to maintain reliable learning signals. This reinforces the method’s value for constructing information-rich and scalable graphs suitable for robust GNN-based modeling.

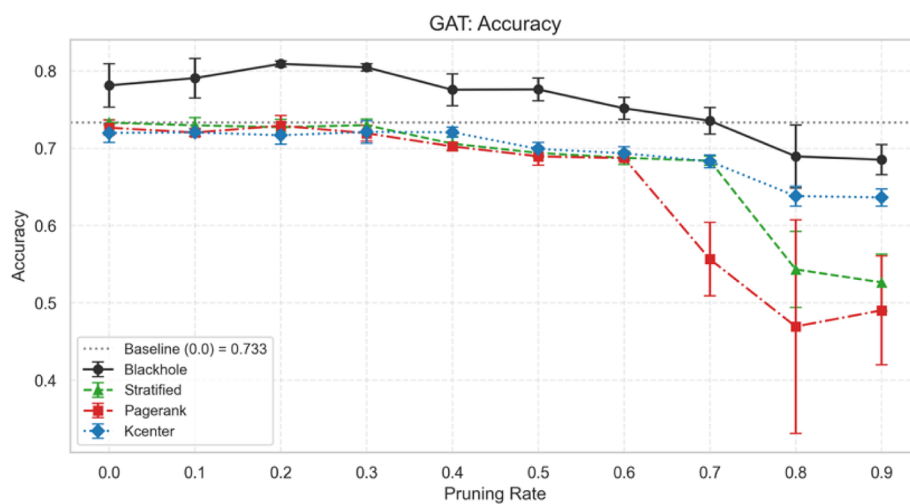
While overall classification accuracy offers a general view of model performance, it does not reveal how effectively each model distinguishes between specific pore-limiting diameter (PLD) classes. To gain deeper insights into class-level performance, we examined confusion matrices for GraphSAGE, GAT, and GCN on the original graph and three selected sparsification thresholds: 0.2, 0.5, and 0.8. These matrices highlight how predictions are distributed across true PLD categories, revealing misclassification patterns, class sensitivities, and structural effects of sparsification.

In Figure 9, which shows the models’ performance on the original (unsparsified) graph, all three GNNs demonstrate clear diagonal dominance, indicating reliable predictions across classes. GraphSAGE achieves particularly strong separation, especially for the “Medium” and “Large” classes. GAT shows modest confusion between adjacent classes, while GCN performs the weakest, with more frequent misclassifications, especially between “Small” and “Medium.”

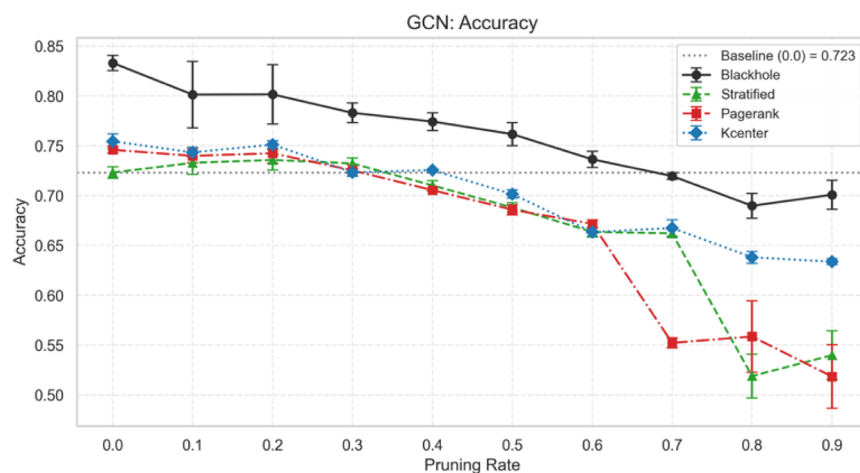
Figure 10 presents confusion matrices at threshold 0.2, where all models retain much of their original classification capability. GraphSAGE remains reliable, although some confusion arises between the “Small” and “Medium” categories. GAT begins to struggle more clearly, with increased off-diagonal predictions, especially between “Medium” and “Large.” GCN maintains moderate performance, though its precision for minority classes appears reduced.

By threshold 0.5 (Figure 11), model degradation becomes prominent. GraphSAGE still provides the most structured output, though predictions for “Small” and “Nonporous” begin to overlap. GAT suffers from diffuse prediction patterns across nearly all categories, while GCN exhibits high confusion and a lack of reliable classification, suggesting significant information loss due to sparsification.

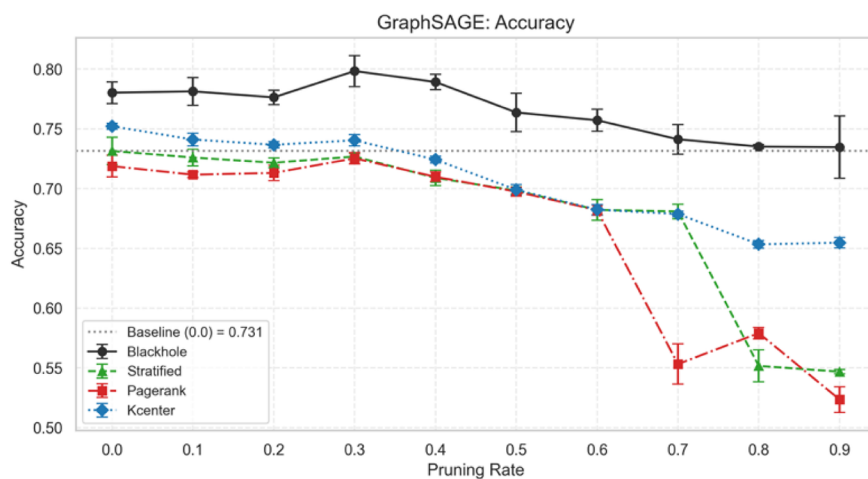
At threshold 0.8 (Figure 12), a partial performance rebound is observed, especially for GraphSAGE, which regains reasonable accuracy for the “Large” and “Nonporous” classes. GAT



(a)



(b)



(c)

Figure 8. Classification accuracy of three GNN models under varying sparsification levels using the Black Hole Strategy and Stratified Sparsification: (a) GAT, (b) GCN, and (c) GraphSAGE. Each point represents the mean accuracy over 10 independent runs, with error bars indicating the standard deviation. Across all models, the Black Hole Strategy consistently yields higher and more stable performance, particularly at moderate to high sparsification levels.

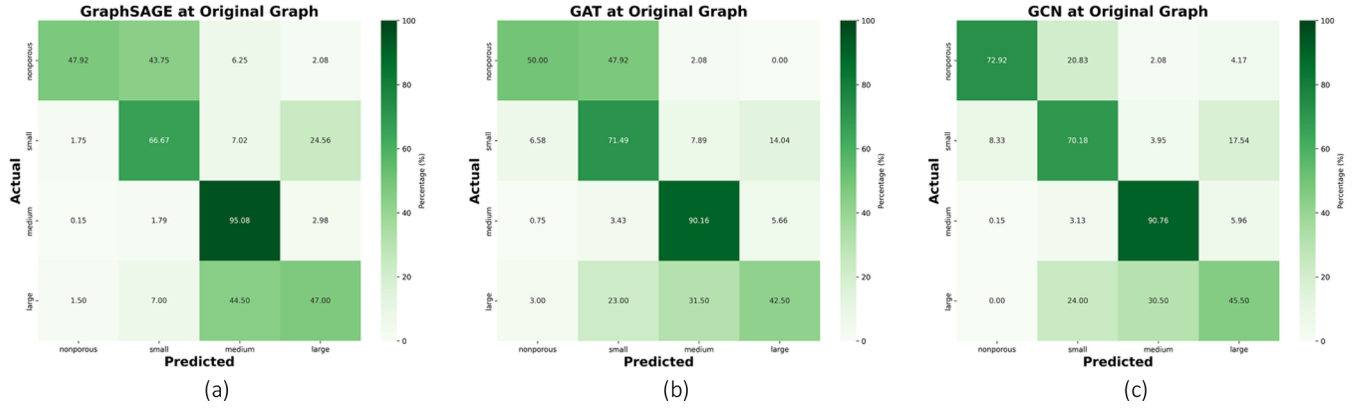


Figure 9. Confusion matrices on the **original (unsparsified)** MOF graph. Panels: (a) GraphSAGE, (b) GAT, (c) GCN.

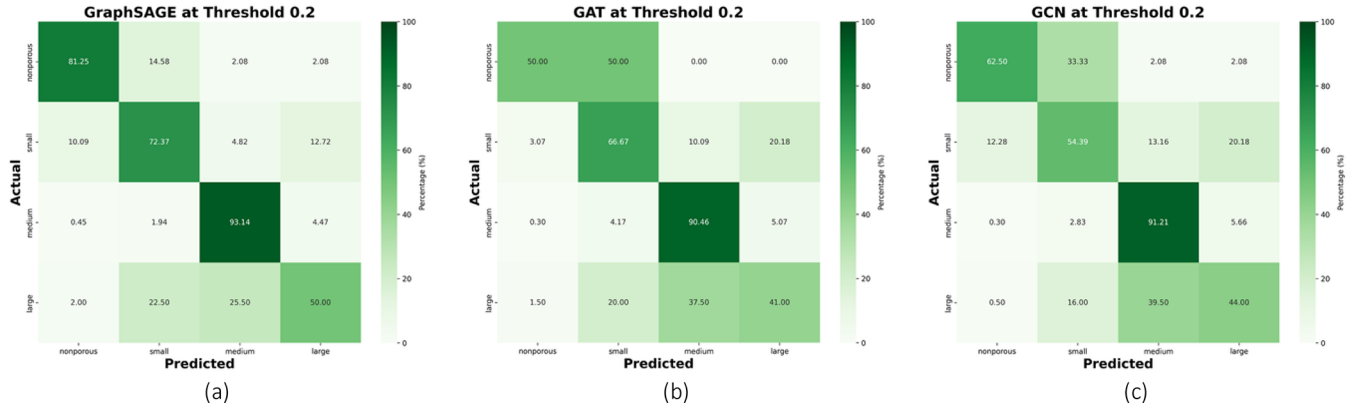


Figure 10. Confusion matrices at sparsification threshold $\tau = 0.2$. Panels: (a) GraphSAGE, (b) GAT, (c) GCN.

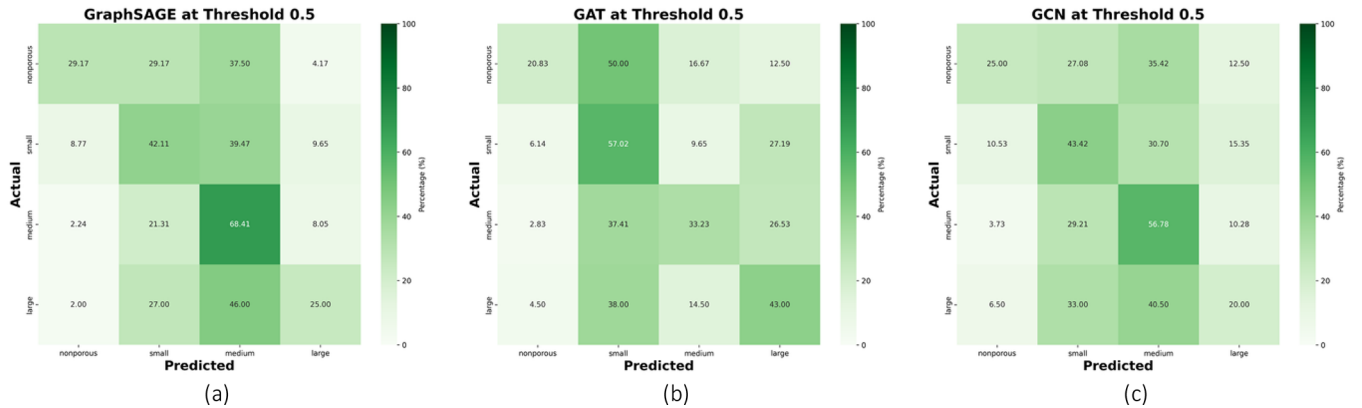


Figure 11. Confusion matrices at sparsification threshold $\tau = 0.5$. Panels: (a) GraphSAGE, (b) GAT, (c) GCN.

continues to show scattered predictions with poor class separation, and GCN remains unstable, highlighting its sensitivity to high sparsity.

Together, these results emphasize that GraphSAGE is more robust under sparsification, while GAT and GCN are more vulnerable to edge and neighborhood degradation. Confusion matrices thus complement average accuracy results by exposing per-class behavior, misclassification trends, and the structural resilience of each GNN model.

3.5. Regression-Based Evaluation of PLD for Domain-Relevant Validation. While classification into PLD categories provides useful insights, pore-limiting diameter (PLD) is inherently continuous, and small variations can significantly

affect adsorption and transport behavior. To capture this nuance, we reformulated the task as a regression problem, predicting PLD values directly. We benchmarked GAT, GCN, and GraphSAGE under four sparsification strategies (Black Hole, Stratified, PageRank, and K-Center) and evaluated regression performance in terms of MAE, RMSE, and R^2 .

As shown in Figures 13–15, the Black Hole Strategy consistently achieves lower error (MAE, RMSE) and higher R^2 across pruning thresholds compared to baselines. For example, at 50% pruning, GAT with BH achieved MAE = 0.836, RMSE = 1.315, $R^2 = 0.683$, whereas Stratified reached only MAE = 1.182, RMSE = 1.781, $R^2 = 0.439$. Similarly, GraphSAGE with BH attained MAE = 0.658, RMSE = 1.050, R^2

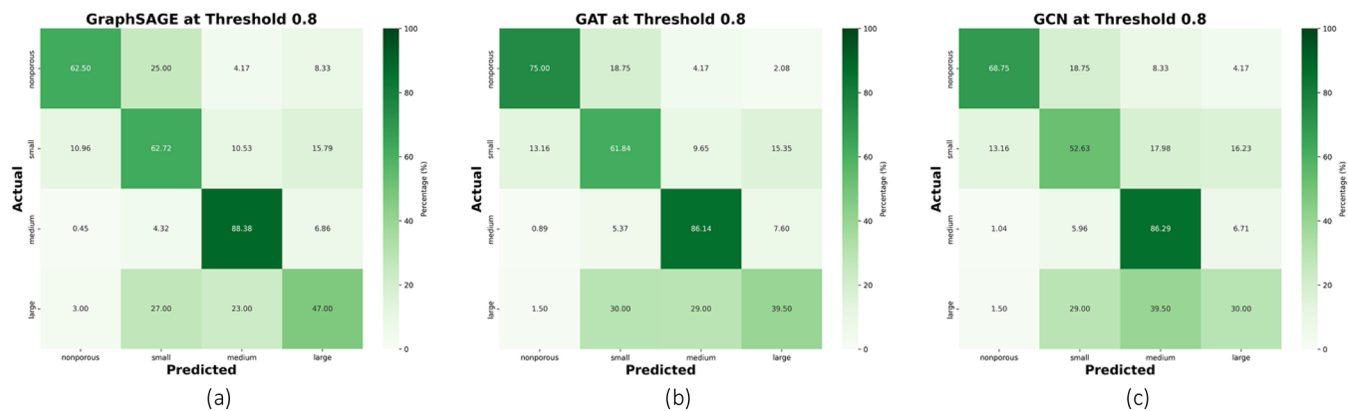


Figure 12. Confusion matrices at sparsification threshold $\tau = 0.8$. Panels: (a) GraphSAGE, (b) GAT, (c) GCN.

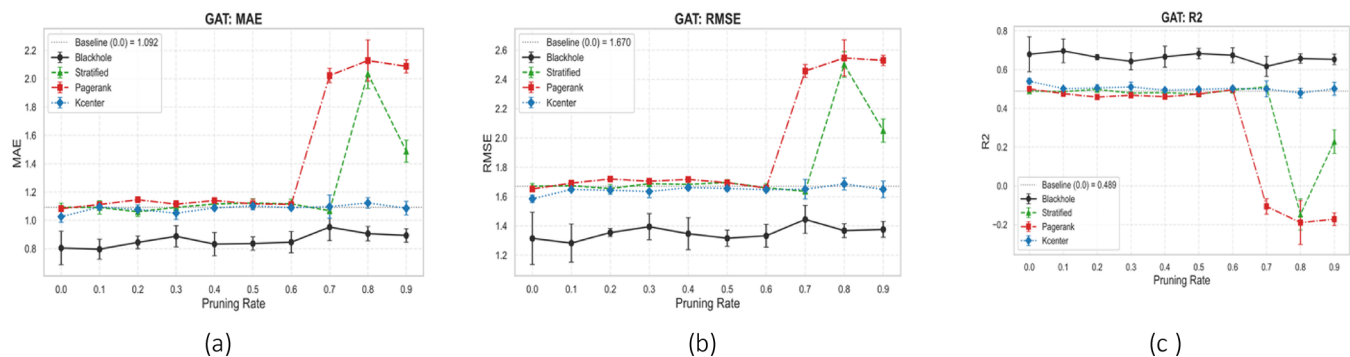


Figure 13. Regression benchmarks for GAT models across pruning thresholds: (a) MAE, (b) RMSE, (c) R^2 .

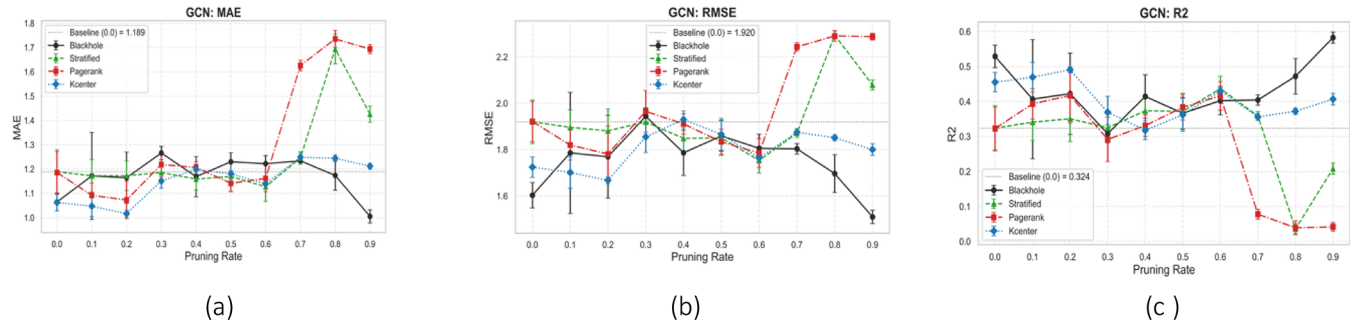


Figure 14. Regression benchmarks for GCN models across pruning thresholds: (a) MAE, (b) RMSE, (c) R^2 .

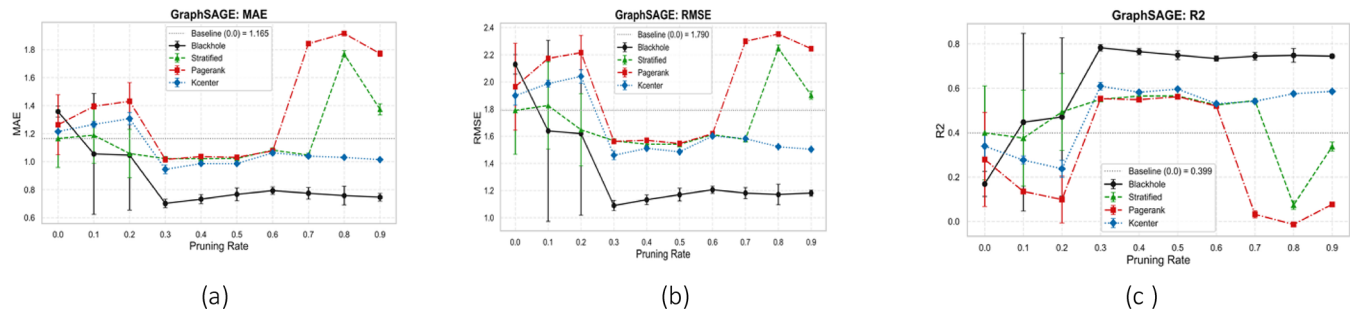


Figure 15. Regression benchmarks for GraphSAGE models across pruning thresholds: (a) MAE, (b) RMSE, (c) R^2 .

= 0.798, representing the strongest regression accuracy overall. These findings confirm that BH-selected subgraphs remain representative not only for classification but also for continuous property prediction, strengthening the method's utility for domain-relevant materials science tasks.

3.6. Topological Evaluation of Black Hole Sparsification. To investigate the topological consequences of sparsification, we compared the Black Hole Strategy against three common baselines—Stratified, PageRank, and K-Center—across several structural metrics. The goal of this analysis was

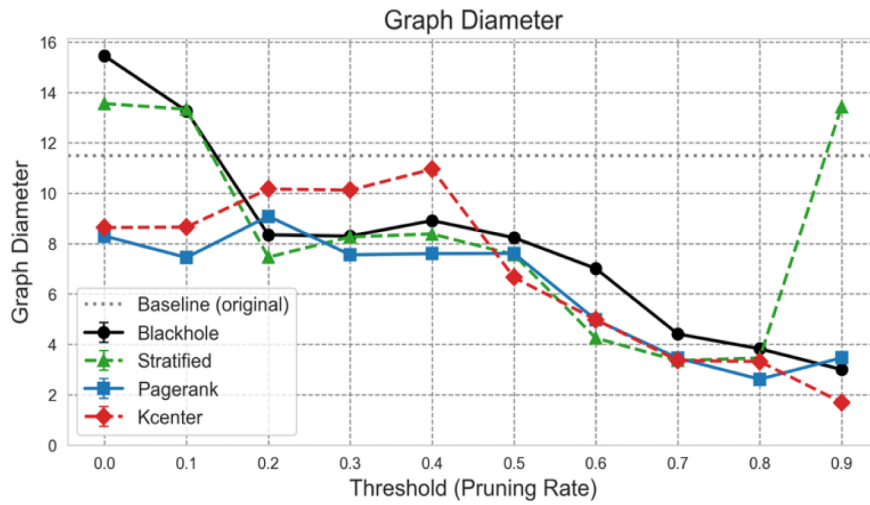


Figure 16. Graph diameter across pruning thresholds. BH preserves long-range paths and global communication, while baselines collapse prematurely.

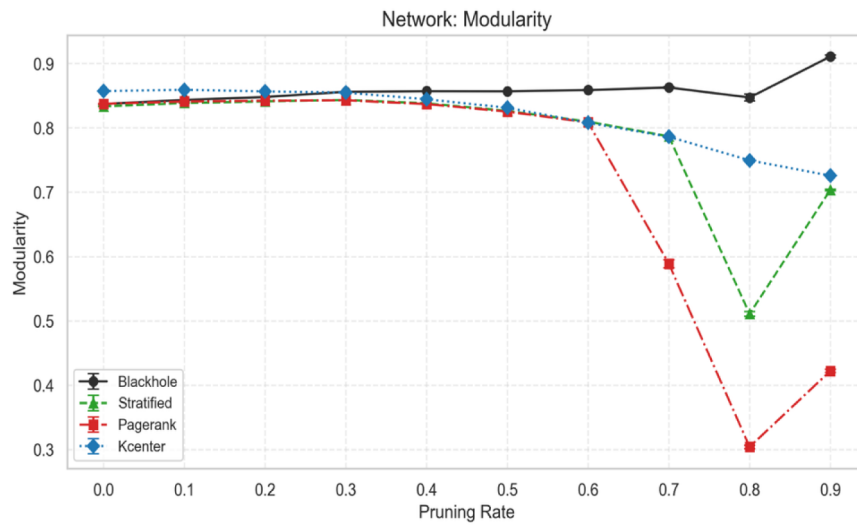


Figure 17. Comparison of modularity across sparsification levels. BH maintains and even enhances community separation, while Stratified and PageRank lose modularity beyond 60% pruning.

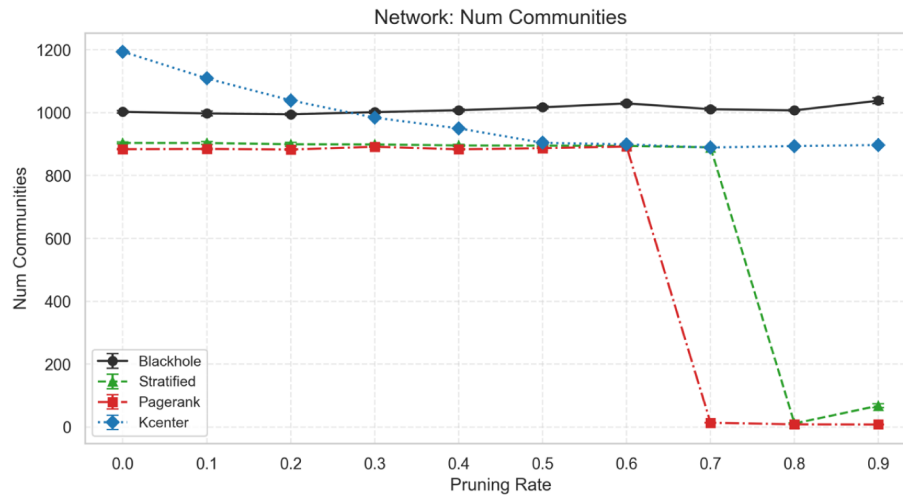


Figure 18. Number of detected communities. BH keeps community counts stable across thresholds, while Stratified and PageRank collapse into fragmented or empty networks.

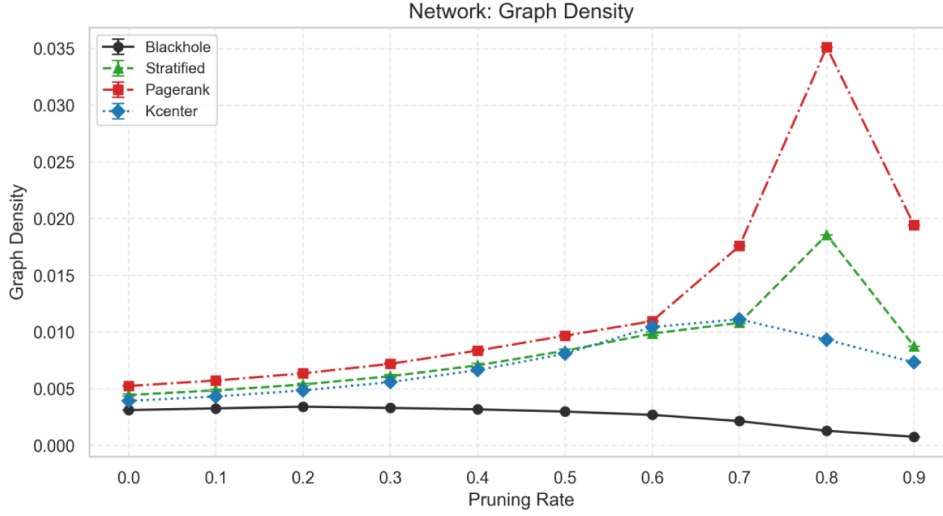


Figure 19. Graph density evolution. BH maintains stable density, avoiding artificial overconcentration, while baselines inflate density under strong pruning.

to determine whether BH preserves both local cohesiveness and global connectivity more effectively than competing approaches. Figures 16–19 summarize the results.

3.6.1. Graph Diameter. Graph diameter captures the longest shortest path within the network and reflects its global reach. As illustrated in Figure 16, BH maintains a broader diameter across all pruning thresholds compared to baselines. This means that sparsified graphs under BH still allow information to travel over long distances, preserving structural integrity. In contrast, Stratified and K-Center rapidly lose diameter after 40% pruning, producing overly compact graphs. PageRank compresses the diameter even more strongly, leading to networks where long-range relationships are almost completely lost. By selectively pruning weaker connections while retaining structurally dominant nodes, BH ensures that sparsification does not compromise global communication.

3.6.2. Graph Modularity. Community structure is another key aspect of network integrity. Figure 17 shows that BH consistently sustains high modularity across all pruning thresholds, even slightly improving community separation at higher levels of pruning. This behavior indicates that the communities detected in the network remain well-defined and interpretable even under strong sparsification. By contrast, the Stratified and PageRank methods collapse dramatically after 60% pruning, losing most community structure and yielding fragmented or artificially blended clusters. K-Center maintains moderate modularity but exhibits a decline at high sparsity. These results demonstrate that BH not only preserves but also enhances modular clarity, a property crucial for downstream clustering and stratified learning tasks.

3.6.3. Number of Communities. The preservation of the number of communities provides further evidence of structural stability. In Figure 18, BH maintains a nearly constant number of communities across all thresholds, indicating that pruning does not eliminate entire structural groups. In contrast, the Stratified and PageRank approaches collapse after 70% pruning, reducing the network to only a handful of communities or even none. K-Center initially inflates the community count—likely due to artificial fragmentation—but eventually collapses as well. The ability of BH to retain stable community counts confirms its effectiveness in balancing pruning with diversity preservation,

ensuring that sparsified graphs remain representative of the original dataset.

3.6.4. Graph Density. Finally, we analyzed graph density, which reflects the ratio of existing edges to the total number of possible connections. As shown in Figure 19, BH maintains stable density values across pruning thresholds, remaining close to the baseline. This stability means that BH does not introduce artificial overconcentration or collapse. By contrast, Stratified, PageRank, and K-Center all exhibit strong density spikes at pruning levels beyond 70%, indicating distorted connectivity where remaining nodes become excessively interconnected. Such artificial densification reduces interpretability and introduces biases into network analysis and GNN training. BH avoids these pitfalls by preserving sparsity in a controlled, realistic manner.

Taken together, these results confirm that the Black Hole Strategy provides a principled and robust framework for sparsification. Unlike the baselines, which either fragment communities, collapse diameters, or artificially inflate density, BH maintains both local structures (community integrity and modularity) and global features (diameter and connectivity). The stability observed across all topological metrics ensures that BH produces graphs that are lightweight, interpretable, and structurally faithful to the original dataset, thereby enabling efficient yet reliable learning of graph neural networks.

3.7. Computational Efficiency and Complexity Analysis. To quantitatively validate the “frugal” characteristics of the Black Hole Strategy, we compared its performance with three general sparsification techniques—Stratified, PageRank, and K-Center—across graph construction duration, peak memory consumption, and model training duration. The results are encapsulated in Table 3, with detailed plots provided in Section S7. Black Hole consistently exhibited the lowest memory footprint (~114 MB at 50% pruning) compared to Stratified (~823 MB) and PageRank (~631 MB), while also cutting training time by nearly half. For example, in GAT, training time was reduced from ~34 s (Stratified) and 49 s (PageRank) to 18 s with Black Hole. K-Center achieved competitive training times (e.g., 28 s in GAT, 1.9 s in GCN), but this benefit was offset by significantly higher construction overhead (75 s vs ~27 s for Black Hole). Overall, Black Hole delivered the best balance,

Table 3. Average Graph Construction Time, Training Time, and Peak Memory Usage at 50% Pruning across Different Sparsification Methods

Model	Method	Graph Construction Time (s)	Training Time (s)	Peak Memory (MB)
GAT	Black Hole	26.7	18.3	114
	Stratified	37.6	33.7	823
	PageRank	33.7	48.5	631
	K-Center	75.4	27.9	193
GCN	Black Hole	26.8	2.7	114
	Stratified	37.6	3.7	823
	PageRank	33.7	4.9	631
	K-Center	75.4	1.9	193
GraphSAGE	Black Hole	26.8	3.1	114
	Stratified	37.6	4.1	823
	PageRank	33.7	5.2	631
	K-Center	75.4	2.2	193

combining fast convergence and minimal memory usage with only moderate graph construction cost.

Theoretically, the intricacy of these methods reinforces the apparent trade-offs in their efficiency. The Black Hole Strategy, characterized by Louvain community detection and centrality calculations, functions with a practical complexity of approximately $O(N \cdot E)$, but community partitioning significantly decreases runtime in practice. PageRank depends on iterative power updates with a complexity of approximately $O(k \cdot E)$, where k is the number of iterations, which is generally minimal. Stratified sampling is computationally efficient, with a time complexity of $O(N)$, as it involves random selection within designated property bins. K-Center, however, is the most costly, necessitating $O(N^2)$ distance evaluations in its greedy implementation (or $O(N \log N)$ with approximations). Collectively, our findings suggest that the Black Hole Strategy achieves the optimal equilibrium between empirical efficiency and theoretical scalability, underscoring its suitability for extensive graph learning in materials science.

In addition to computational benchmarks, we also examined the effect of synthetic redundancy to provide direct evidence that low-information nodes hinder model generalization. Redundancy was introduced by duplicating a fraction (r) of nodes with identical features and connections, and performance was evaluated for GAT, GCN, and GraphSAGE with and without the Black Hole Strategy, taking the best accuracy across sparsification thresholds. As summarized in Table 4 and illustrated in Section S7, redundancy consistently degraded performance in baseline models, particularly at higher levels ($r \geq 0.2$). In contrast, BH mitigated this effect by selectively removing low-information nodes, yielding accuracy improvements of up to +9.5% for GAT at $r = 0.0$ –0.1. These results provide direct evidence that redundancy can obscure structure–property relationships while confirming that BH enhances both predictive accuracy and computational efficiency, in line with the principles of frugal machine learning.

4. CONCLUSIONS AND FUTURE DIRECTION

This study introduced the Black Hole Strategy, a scalable and structure-aware sparsification method for curating high-quality subsets from large and redundant materials datasets. By reducing data volume while preserving structural and chemical diversity, the strategy demonstrates that representative, well-structured data can be more valuable than sheer quantity. Both

Table 4. Maximum Classification Accuracy across Sparsification Thresholds for GAT, GCN, and Graphsage under Increasing Synthetic Redundancy Levels (r)

Model	r (Redundancy Level)	Black Hole	Stratified	PageRank	K-Center
GAT	0.0	0.571	0.540	0.564	0.575
	0.1	0.564	0.550	0.572	0.608
	0.2	0.564	0.542	0.566	0.551
	0.3	0.558	0.538	0.572	0.573
GCN	0.0	0.550	0.523	0.525	0.566
	0.1	0.560	0.526	0.555	0.561
	0.2	0.540	0.559	0.548	0.550
	0.3	0.553	0.526	0.538	0.533
GraphSAGE	0.0	0.588	0.550	0.573	0.624
	0.1	0.563	0.529	0.541	0.588
	0.2	0.594	0.542	0.559	0.608
	0.3	0.591	0.560	0.549	0.595

classification and regression experiments confirm its robustness across predictive tasks, showing that the approach maintains strong performance even under significant data reduction. Beyond predictive accuracy, it enhances interpretability, reduces noise, and improves computational efficiency, making it particularly suitable for large-scale screening and machine learning–driven discovery. Importantly, the approach aligns with the goals of the consortium FAIRmat (<https://www.fairmat-nfdi.eu/fairmat/>), which is building a FAIR (Findable, Accessible, Interoperable, Reusable) data infrastructure for condensed-matter physics and chemical physics. By generating optimized datasets, the Black Hole Strategy supports both machine learning applications and knowledge-driven exploration within this framework.

Future research can extend the Black Hole Strategy along several promising directions. Representing MOFs at the atomic scale, rather than as higher-level nodes, could enable more fine-grained property prediction and tighter integration with molecular simulation workflows. The sparsified MOF networks derived from the Black Hole Strategy also provide a solid foundation for advanced applications, including generative modeling, optimization-driven material selection, and retrieval-augmented generation with large language models (LLMs). Linking this strategy with GraphRAG architectures or knowledge graph systems may further facilitate intelligent querying, semantic reasoning, and explainable AI in materials science. Moreover, the Black Hole Strategy offers natural synergies with semisupervised and active learning paradigms. For example, the gravity score could serve as an acquisition function to prioritize labeling in data-scarce environments or as a heuristic to guide label propagation in semisupervised frameworks.

A particularly significant avenue of future work lies in combining the Black Hole Strategy with the Lotus Effect Optimization (LEO) algorithm,^{30,31} both of which were originally proposed by our group. LEO, which employs nature-inspired mechanisms to balance exploration and exploitation, can be adapted as a gravity-based optimization method to complement the Black Hole framework. This hybridization holds strong potential to advance scalable discovery pipelines in materials science, especially by accelerating the identification of optimal MOFs for drug discovery and related biomedical applications. Such integration would also align the Black Hole Strategy with broader trends in data-efficient machine learning, further increasing its utility for

chemical and materials datasets where labeled data remain limited.

AUTHOR INFORMATION

Corresponding Authors

Mehrdad Jalali – Institute of Functional Interfaces (IFG), Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen 76344, Germany; Department of Information, SRH University of Applied Sciences Heidelberg, Heidelberg 69123, Germany; orcid.org/0000-0003-2465-4933; Email: mehrdad.jalali@srh.de

A. D. Dinga Wonanke – Faculty of Chemistry and Food Chemistry, TU Dresden, Dresden 01069, Germany; Email: dinga.wonanke@tu-dresden.de

Authors

Pascal Friederich – Institute of Theoretical Informatics, Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany; Institute of Nanotechnology, Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen 76344, Germany; orcid.org/0000-0003-4465-1465

Christof Wöll – Institute of Functional Interfaces (IFG), Karlsruhe Institute of Technology (KIT), Eggenstein-Leopoldshafen 76344, Germany; orcid.org/0000-0003-1078-3304

Author Contributions

M.J. conceived and supervised the project. A.D.D.W. developed and implemented the methodology. P.F. provided theoretical guidance, contributed to validation and evaluation, and suggested improvements. C.W. contributed to project conceptualization, methodological development, and critical manuscript revision. All authors discussed the results and contributed to the preparation of the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) as part of the FAIRmat project (Project ID: 460197019), which focuses on implementing FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles in materials science. Open-access funding for this research was enabled and organized by Projekt DEAL, ensuring the broader accessibility and dissemination of the findings. A.D.D.W. acknowledges funding from the European Union Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101107360.

REFERENCES

- (1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.
- (2) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **2019**, *5*, 83.
- (3) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066–8129.
- (4) Tsai, M. H.; Yeh, J. W. High-entropy alloys: A critical review. *Mater. Res. Lett.* **2014**, *2* (3), 107–123.
- (5) Furukawa, H.; Cordova, K. E.; O’Keeffe, M.; Yaghi, O. M. chemistry and applications of metal-organic frameworks. *Science* **2013**, *341* (6149), 1230444.
- (6) Li, D.; Yadav, A.; Zhou, H.; Roy, K.; Thanasekaran, P.; Lee, C. Advances and Applications of Metal-Organic Frameworks (MOFs) in Emerging Technologies: A Comprehensive Review. *Global Challenges* **2024**, *8*, 2300244.
- (7) Jia, T.; Gu, Y.; Li, F. Progress and potential of metal-organic frameworks (MOFs) for gas storage and separation: A review. *J. Environ. Chem. Eng.* **2022**, *10* (5), 108300.
- (8) Nguyen, L. H. T.; Mirzaei, A.; Kim, J. Y.; Phan, T. B.; Dai Tran, L.; Wu, K. C. W.; Kim, H. W.; Kim, S. S.; Doan, T. L. H. Advancements in MOF-based resistive gas sensors: synthesis methods and applications for toxic gas detection. *RSC Adv.* **2025**, *10*, 1025–1053.
- (9) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. *Acta Crystallogr., Sect. B: struct. Sci., Cryst. Eng. Mater.* **2016**, *72* (2), 171–179.
- (10) Moghadam, P. Z.; Li, A.; Wiggan, S. B.; Tao, A.; Maloney, A. G. P.; Wood, P. A.; Ward, S. C.; Fairen-Jimenez, D.; et al. Development of a Cambridge Structural Database Subset: A Collection of Metal-Organic Frameworks for Past, Present, and Future. *Chem. Mater.* **2017**, *29*, 2618–2625.
- (11) Burner, J.; Luo, J.; White, A.; Mirmiran, A.; Kwon, O.; Boyd, P. G.; Maley, S.; Gibaldi, M.; Simrod, S.; Ogden, V.; Woo, T. K.; et al. ARC-MOF: A Diverse Database of Metal-Organic Frameworks with DFT-Derived Partial Atomic Charges and Descriptors for Machine Learning. *Chem. Mater.* **2023**, *35* (3), 900–916.
- (12) Wang, Y.; Jiang, Z. J.; Wang, D. R.; Lu, W.; Li, D. Machine Learning-Assisted Discovery of Propane-Selective Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2024**, *146* (10), 6955–6961.
- (13) Bangian, E.; Jalali, M.; Houshmand, M. Inverse Link Prediction with Graph Convolutional Networks for Knowledge-Preserving Sparsification in Cheminformatics. *J. Big Data* **2025**, *12*, 176.
- (14) Jalali, M.; Wonanke, A. D. D.; Wöll, C. MOFGalaxyNet: a social network analysis for predicting guest accessibility in metal-organic frameworks utilizing graph convolutional networks. *J. Cheminf.* **2023**, *15* (1), 94.
- (15) Jalali, M.; Tsotsalas, M.; Wöll, C. MOFSocialNet: Exploiting Metal-Organic Framework Relationships via Social Network Analysis. *Nanomaterials* **2022**, *12* (4), 704.
- (16) Shayanmehr, M.; Aarabi, S.; Ghaemi, A.; Hemmati, A. A data driven machine learning approach for predicting and optimizing sulfur

compound adsorption on metal organic frameworks. *Sci. Rep.* **2025**, *15* (1), 3138.

(17) Colón, Y. J.; Snurr, R. Q. High-throughput computational screening of metal-organic frameworks. *Chem. Soc. Rev.* **2014**, *43*, 5735–5749.

(18) Chung, Y. G.; Haldoupis, E.; Bucior, B. J.; Haranczyk, M.; Lee, S.; Zhang, H.; Vogiatzis, K. D.; Milisavljevic, M.; Ling, S.; Camp, J. S.; et al. Advances, Updates, and Analytics for the Computation-Ready, Experimental Metal-Organic Framework Database: CoRE MOF 2019. *J. Chem. Eng. Data* **2019**, *64* (12), 5985–5998.

(19) Mukherjee, K.; Colón, Y. J. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks. *Mol. Simul.* **2021**, *47* (10–11), 857–877.

(20) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **2024**, *6* (2), 161–169.

(21) Choudhary, K.; Yildirim, T.; Siderius, D. W.; Kusne, A. G.; McDannald, A.; Ortiz-Montalvo, D. L. Graph neural network predictions of metal organic framework CO₂ adsorption properties. *Comput. Mater. Sci.* **2022**, *210*, 111388.

(22) Ruff, R.; Reiser, P.; Stühmer, J.; Friederich, P. Connectivity optimized nested line graph networks for crystal structures. *Digital Discovery* **2024**, *3* (3), 594–601.

(23) Chen, L. Y.; Hsu, T. W.; Hsiung, T. C.; Li, Y. P. Deep Learning-Based Increment Theory for Formation Enthalpy Predictions. *J. Phys. Chem. A* **2022**, *126* (41), 7548–7556.

(24) Muthiah, B.; Li, S. C.; Li, Y. P. Developing machine learning models for accurate prediction of radiative efficiency of greenhouse gases. *J. Taiwan Inst. Chem. Eng.* **2023**, *151*, 105123.

(25) Ghouchan Nezhad Noor Nia, R.; Jalali, M.; Mail, M.; Ivanisenko, Y.; Kübel, C. Machine Learning Approach to Community Detection in a High-Entropy Alloy Interaction Network. *ACS Omega* **2022**, *7* (15), 12978–12992.

(26) Ghouchan Nezhad Noor Nia, R.; Jalali, M.; Houshmand, M. A Graph-Based k-Nearest Neighbor (KNN) Approach for Predicting Phases in High-Entropy Alloys. *Appl. Sci.* **2022**, *12* (16), 8021.

(27) Lang, K.; Liberty, E.; Shmakov, K. Stratified sampling meets machine learning In *33rd International Conference on Machine Learning, ICML 2016*; PMLR, 2016.

(28) Leskovec, J.; Faloutsos, C. Sampling from large graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM, 2006. DOI: .

(29) Bateni, M. H.; Esfandiari, H.; Fischer, M.; Mirrokni, V. Extreme k-Center Clustering. In *35th AAAI Conference on Artificial Intelligence, AAAI 2021*; PKP, 2021. DOI: .

(30) Dalirinia, E.; Jalali, M.; Yaghoobi, M.; Tabatabaee, H. Lotus effect optimization algorithm (LEA): a lotus nature-inspired algorithm for engineering design optimization. *J. Supercomput.* **2024**, *80* (1), 761–799.

(31) Dalirinia, E.; Yaghoobi, M.; Tabatabaee, H.; Chandna, S.; Jalali, M. Multimodal Lotus Effect Algorithm for Engineering Optimization Problems. *Eng. Rep.* **2025**, *7* (4), No. e70137.