Exploring Heterophily in Graph-level Tasks

Qinhan Hou[1,2], **Yilun Zheng**[3], **Xichun Zhang**[2], **Sitao Luan**[4,5,*], **Jing Tang**[2,*]
[1]Doctoral Program of Computer Science, University of Helsinki, Helsinki, Finland
[2]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[3]Centre for Information Sciences and Systems, Nanyang Technological University, Singapore
[4]University of Montreal, Canada
[5]Mila, Quebec Artificial Intelligence Institute, Canada
[*]Corresponding author, main supervision

Abstract

While heterophily has been widely studied in node-level tasks, its impact on graph-level tasks remains unclear. We present the first analysis of heterophily in graph-level learning, combining theoretical insights with empirical validation. We first introduce a taxonomy of graph-level labeling schemes, and focus on motif-based tasks within local structure labeling, which is a popular labeling scheme. Using energy-based gradient flow analysis, we reveal a key insight: unlike frequency-dominated regimes in node-level tasks, motif detection requires mixed-frequency dynamics to remain flexible across multiple spectral components. Our theory shows that motif objectives are inherently misaligned with global frequency dominance, demanding distinct architectural considerations. Experiments on synthetic datasets with controlled heterophily and real-world molecular property prediction support our findings, showing that frequency-adaptive model outperform frequency-dominated models. This work establishes a new theoretical understanding of heterophily in graph-level learning and offers guidance for designing effective GNN architectures.

1 Introduction

Graph Neural Networks (GNNs) have achieved success in learning from graph-structured data, demonstrating strong performance across diverse domains including social networks [19, 26] and molecular property prediction [41, 45]. Many popular GNN architectures, such as Graph Convolutional Networks (GCNs) [11], are designed under the homophily assumption, *i.e.,* connected nodes tend to share similar features or labels [29, 14, 25]. However, many real-world graphs exhibit heterophily, where neighboring nodes have dissimilar characteristics [47, 24]. While the challenges posed by heterophily have been extensively studied on node-level tasks [23, 21, 22, 44, 43, 42], its impact on graph-level tasks remains poorly understood.

This work is the first to study heterophily in graph-level tasks. We introduce a taxonomy that classifies such tasks into three types by their labeling mechanisms, focusing on motif-based tasks where labels depend on discriminative subgraphs (*motifs*). From an energy and gradient flow perspective [12], our analysis shows that graph-level tasks have distinct frequency preferences from node-level tasks, as motif detection misaligns with the global nature of low- and high-frequency dominant regimes. This misalignment challenges the effectiveness of GNN under heterophily settings. We provide both theoretical and empirical evidence, offering new insights into heterophily's role in graph-level prediction and guiding the design of more adaptive GNNs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Graph Machine Learning.

2 Task Taxonomy, Notations and Background

2.1 Task Taxonomy

Compared to node-level tasks, the labeling schemes of graph-level tasks make it challenging to establish a simple and general relation between graph labels and certain graph properties. To enable a simplified and systematic discussion, we categorize the labeling schemes into three main types based on the following criterion derived from practical applications.

Aggregated Node Features. In this scenario, graph labels are primarily determined by aggregated node features. For instance, a graph may be assigned to a particular class if the average value of a specific node feature across all nodes exceeds a given threshold, or if a certain proportion of nodes belong to a particular latent class, *e.g.,* online community detection based on aggregated user demographics [30, 9, 5].

Local Structure. Labels depend on local structural patterns and node-level features. For example, a label may be assigned based on the presence of a specified number of particular motifs (*e.g.,* triangles) within the graph. These motifs may predominantly consist of either homophilic or heterophilic nodes, *e.g.,* molecular classification based on the presence of specific pharmacophores or toxic substructures [7, 40, 18, 36].

Global Structure. In this case, labels are determined by global structural properties of the graph, such as its diameter or overall connectivity. The label thus reflects a purely structural characteristic of the graph, *e.g.,* metabolic network categorization of different organisms based on global connectivity patterns, such as scale-free vs. random network topologies [46, 28, 33].

To give a more intuitive understanding, we list some real-world applications according to these three types of tasks in the Appendix A. Note that the above three categories are not exclusive, and a graph can be classified by mixture criteria. To simplify the discussion, we focus on the local structure labeling in this work, which is common in real world.

2.2 Energy-Based Framework for Understanding GNN Dynamics

Recent work by Di Giovanni et al. [12] provides a rigorous framework for analyzing GNNs as dynamical systems that minimize a generalized energy functional. This framework reveals that under certain conditions, the training dynamic of GNN will lead to a global frequency-dominated regime, leading to a bipolar convergence of node features. We review this framework briefly.

GNN Dynamics as Gradient Flow Consider an undirected and connected graph $G = (V, E)$, where nodes $v = \{v_1, v_2, \ldots, v_n\} \in V$ have features $\{\mathbf{f}_i \in \mathbb{R}^d : v_i \in V\}$, and edge set denotes $E \subset V \times V$. The feature matrix $\mathbf{F} \in \mathbb{R}^{n \times d}$ consists of $\mathbf{f}_i$ as its rows. According to [3], Message Passing Neural Networks (MPNNs) [11] update the layer $t + 1$ via:

$$\dot{\mathbf{F}} = \mathbf{F}(t+1) - \mathbf{F}(t) = \sigma(-\mathbf{F}(t)\mathbf{\Omega}_t + \mathbf{A}\mathbf{F}(t)\mathbf{W}_t - \mathbf{F}(0)\tilde{\mathbf{W}}_t) \tag{1}$$

where Ω_t, $\mathbf{W}_t$, and $\tilde{\mathbf{W}}_t$ are learnable matrices performing feature transformations, σ is the non-linear activation function, and $\mathbf{A}$ is the adjacency matrix which aggregates neighbor information.

A **gradient flow** is a special dynamical system which is defined by an ordinary differential equation $\dot{\mathbf{F}}(t) = -\nabla\mathcal{E}(\mathbf{F}(t))$. The dynamic in Eq.1 corresponds to a gradient flow of the energy functional:

$$\mathcal{E}_\theta(\mathbf{F}) \;=\; \sum_i \langle \mathbf{f}_i, \mathbf{\Omega}\mathbf{f}_i \rangle \;-\; \sum_{i,j} A_{ij} \langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j \rangle \;+\; \varphi^0(\mathbf{F}, \mathbf{F}(0)) \tag{2}$$

given the conditions that the weight matrices $\mathbf{\Omega}$ and $\mathbf{W}$ are symmetric [1], and φ^0 is a pre-defined function to calculate the distance between $\mathbf{F}$ and the source $\mathbf{F}(0)$. Note that the Dirichlet energy $\mathcal{E}^{Dir}(\mathbf{F}(t)) := \frac{1}{2}\sum_{(i,j)\in E} \|\nabla\mathbf{F}(t)\|^2$ is a special case of $\mathcal{E}_{\mathbf{\Theta}}$ when $\mathbf{\Omega} = \mathbf{W} = \mathbf{I}_d$ and $\varphi^0 = 0$.

[1] Note that the symmetric is due to the result of derivative of the functional Eq. 1, not the requirement of the matrices.

Asymptotic Frequency-Dominated Regimes This energy framework reveals that linear GNNs converge to one of two asymptotic behaviors, characterized by the relationship between the graph Laplacian spectrum and the eigenvalues of the weight matrix $\mathbf{W}$. Let $\mathbf{\Delta} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ be the normalized Laplacian with ordered eigenvalues $0 = \boldsymbol{\lambda}_0 \leq \cdots \leq \boldsymbol{\lambda}_{n-1}$ and corresponding eigenvectors $\{\boldsymbol{\phi}_0, \ldots, \boldsymbol{\phi}_{n-1}\}$. We give the symmetric definition for the frequency-dominated dynamics and a theorem to decide the dynamics in a simplified version of MPNN.

Definition 1 (Frequency-Dominated Dynamics)**.** *The dynamics of a GNN depends on the limiting behavior of the normalized Dirichlet energy $\mathcal{E}^{Dir}(\mathbf{F}(t))/\|\mathbf{F}(t)\|^2$. If it converges to the eigenvalue $\boldsymbol{\lambda}_0$, we call the dynamic Low-Frequency-Dominant (LFD). Conversely, if it converges to the eigenvalue $\boldsymbol{\lambda}_{n-1}$, we call it High-Frequency-Dominant (HFD).*

We refer a lemma to illustrate the condition to decide whether the MPNN is HFD or LFD.

Lemma 1 (Theorem 4.3 in [12])**.** *Given a continuous MPNN of the form $\dot{\boldsymbol{F}}(t) = \boldsymbol{A}\boldsymbol{F}(t)\boldsymbol{W}$, let $\boldsymbol{\mu}_0 \leq \boldsymbol{\mu}_1 \leq \cdots \leq \boldsymbol{\mu}_{d-1}$ be the eigenvalues of $\boldsymbol{W}$. If $|\boldsymbol{\mu}_0|(\boldsymbol{\lambda}_{n-1} - 1) > \boldsymbol{\mu}_{d-1}$, then for almost every $\boldsymbol{F}(0)$, the MPNN is HFD. Conversely, if $|\boldsymbol{\mu}_0|(\boldsymbol{\lambda}_{n-1} - 1) < \boldsymbol{\mu}_{d-1}$, then for almost every input $\boldsymbol{F}(0)$, the MPNN is LFD.*

From the lemma, the dynamic of the network depends on the biggest and lowest eigenvalues of $\mathbf{W}$: $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_{d-1}$. Since the network is a gradient flow along the energy functional described in Eq. 2, the network is trained to minimize the energy. We will discuss below how will this energy decreasing affect $\mathbf{W}$ and its eigenvalues, especially under heterophily situation.

2.3 Graph Heterophily in Energy-based Framework

Heterophily refers to the tendency of connected nodes to have dissimilar features or labels, in contrast to homophily where neighboring nodes are similar [47, 22]. Within the energy functional framework, heterophily has a direct correspondence to the spectral behavior of GNNs.

Recall the weight matrix $\mathbf{W}$ in Eq. 2, it can be rewritten as $\mathbf{W} = \mathbf{\Theta}_+^\top\mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top\mathbf{\Theta}_-$ by decomposing it into components with positive and negative eigenvalues (see the derivation at D.1) . The pairwise interaction term $\sum_{i,j} A_{ij}\langle \mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle$ in Eq. 2 captures the relationship between connected nodes. When $\mathbf{W}$ has predominant positive eigenvalues, which leads to $[\mathbf{\Theta}_+^\top\mathbf{\Theta}_+]_{i,j} \gg [\mathbf{\Theta}_-^\top\mathbf{\Theta}_-]_{i,j}$, the maximization of the pairwise interaction term will encourage neighboring nodes to have aligned representations. This naturally smooths the features across connected nodes, promotes GNN to LFD dynamics, so that benefits GNN performance on homophilic graphs.

Conversely, when $\mathbf{W}$ has significant negative eigenvalues (*i.e.,* $\mathbf{\Theta}_-^\top\mathbf{\Theta}_-$ dominates), the optimization of the pairwise interaction term will encourage connected nodes to have *anti-aligned* or dissimilar representations. This will drive the system toward HFD dynamics, where high-frequency components dominate, and sharp differences emerge between connected nodes. And the GNN will end up with an energy landscape that favors heterophilic patterns.

The eigenvalue decomposition $\mathbf{W} = \mathbf{\Theta}_+^\top\mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top\mathbf{\Theta}_-$ thus reveals a fundamental trade-off: the energy framework naturally biases GNNs toward either global homophily (LFD) or global heterophily (HFD), but not both simultaneously . This global preference creates challenges for tasks that require *local adaptation*—where some regions of the graph exhibit homophilic patterns (*e.g.,* within motif instances) while others exhibit heterophilic patterns (*e.g.,* at motif boundaries) .

3 Motif Detection Requires Mixed-Frequency Dynamics

Given that the energy framework naturally leads GNNs to frequency-dominated regimes, in this section, we will show that graph-level motif detection represents a fundamentally different class of problems which cannot be solved by either pure low-frequency or high-frequency dynamics.

3.1 Shift Motif Detection to Node-level Task

A given motif $M = (V_M, E_M)$ is a connected graph pattern, where $V_M \subseteq V$ and $E_M = \{(u, v) \in E : u, v \in V_M\}$. There exists a graph isomorphism $\psi : V_M \to V'$ such that $(u, v) \in E_M$ if and only if $(\psi(u), \psi(v)) \in E'$. While often framed as a graph-level problem (*i.e.,* determining if a graph

contains a motif), its inherent dependency on local structure makes it easy to be formulated as a node-level problem. This perspective can shift the objective from a graph-level task to a node-level task, which is to identify all the nodes that belong to the motif. We define the objective of the motif detection task as a node-level task:

Definition 2 (Node-Level Motif Detection)**.** *For a graph $G = (V, E)$ and motif M, we assign binary labels $y_i \in \{0, 1\}$ to each node $i \in V$, where $y_i = 1$ if the node i is part of any subgraph isomorphic to M, and 0 otherwise. The task is to learn $f_{node} : G \to \{0, 1\}^{|V|}$ that predicts $\mathbf{y} = (y_1, \ldots, y_{|V|})$ using a GNN encoder E_{node} and classifier.*

Furthermore, we prove the equivalence of the task objectives in different levels.

Proposition 1 (Equivalence of Node-Level and Graph-Level Motif Detection.)**.** *For graph $G = (V, E)$ and motif M, the following objectives are equivalent: (i) detect if G contains a subgraph isomorphic to M; (ii) detect whether $\exists i \in V$ such that $y_i = 1$ in the node-level task.*

Heterophily Patterns in Motif Detection The node-level motif detection task reveals a fundamental heterophily challenge that distinguishes graph-level tasks from node-level tasks. Unlike traditional node classification where heterophily is characterized globally across the entire graph, motif detection requires handling *spatially-varying* heterophily patterns.

Effective motif detection demands three distinct connectivity behaviors. First, **intra-motif homophily** requires nodes within the same motif instance to have similar representations for consistent detection ($y_i = y_j = 1$ for $i, j \in V_M$). Second, **inter-motif heterophily** necessitates strong representational boundaries between motif participants and non-participants ($\mathbf{f}_i \neq \mathbf{f}_j$ for $i \in V_M, j \notin V_M$ when $(i, j) \in E$). Finally, **context-dependent adaptation** means the same edge (i, j) may require homophilic smoothing if both nodes are motif participants, or heterophilic sharpening if they represent a motif-background boundary.

This spatial heterogeneity in connectivity requirements creates a fundamental mismatch with the energy framework's global frequency preferences, as we demonstrate below.

3.2 Theoretical Incompatibility with Frequency-Dominated Regimes

From Heterophily to Mixed-Frequency Requirements The spatially-varying heterophily patterns required for motif detection directly translate to mixed-frequency requirements in the spectral domain. Recall from Section 2.3 that LFD dynamics correspond to global homophily (feature smoothing), while HFD dynamics correspond to global heterophily (feature sharpening). However, motif detection requires *both* behaviors simultaneously but at different spatial locations.

Specifically, the optimal node representation $\mathbf{f}_i^*$ for motif detection must satisfy conflicting spectral requirements: (**i**) low-frequency components are needed within motif instances to maintain intra-motif consistency ($\mathbf{f}_i^* \approx \mathbf{f}_j^*$ for $i, j \in V_M$), (**ii**) high-frequency components are essential at motif boundaries to create discriminative separation ($\|\mathbf{f}_i^* - \mathbf{f}_j^*\|$ large for $i \in V_M, j \notin V_M$), and (**iii**) medium-frequency components may be required for motifs of specific structural scales.

The energy functional's global optimization toward either LFD or HFD regimes cannot accommodate this spatial heterogeneity. A purely LFD approach would blur motif boundaries through over-smoothing, while a purely HFD approach would fragment intra-motif coherence through excessive sharpening. This fundamental incompatibility explains why frequency-dominated GNNs struggle with motif detection across different heterophily settings.

To give a theoretical analysis on the incompatibility with the frequency-dominated regimes, we first draw a lemma that the performance of GNN on motif-based graph-level tasks is upper bounded by its performance on the node-level task defined in Def. 2.

Lemma 2. *The node-level motif detection function $f_{node} : \mathcal{G} \to \{0, 1\}^{|V|}$ contains sufficient information to solve the graph-level motif detection problem. Specifically, motif M exists in graph G if and only if $\|\mathbf{y}\|_0 > 0$, where $\mathbf{y} = f_{node}(G)$ and $\|\cdot\|_0$ denotes the ℓ_0 norm.*

We then show (informally) below that, if there exists an ideal encoder $E_{\text{node}} : \{\mathbf{f}_i^{(0)}\}_{i=1}^{|V|} \to \{\mathbf{f}_i^{(t)}\}_{i=1}^{|V|}$ for node-level motif detection, it is not aligned with the frequency-dominated dynamics.

Theorem 1. *The frequency-dominated regimes are suboptimal for node-level motif detection tasks.*

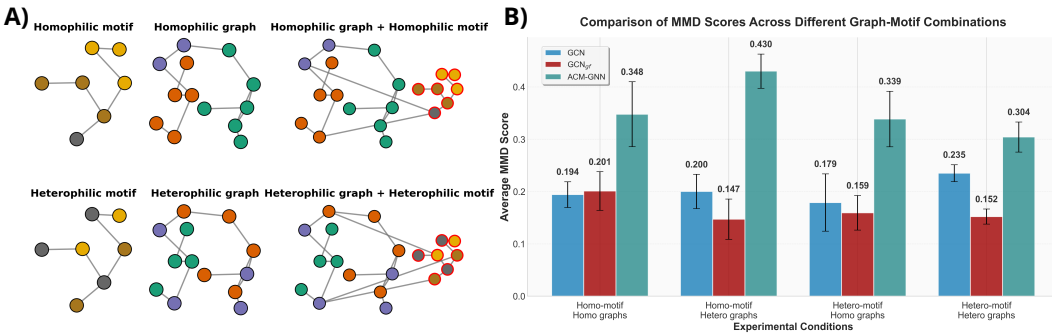

Figure 1: **(A)** The motifs and background graphs in two different settings (homophily and heterophily). Node colors represent different node features/labels. The example illustrates motifs and backbones with distinct feature distributions. **(B)** MMD scores of different conditions. The MMD scores are calculated on the test sets between graphs with and without motifs attached.

The proofs for the above results are given in the Appendix E. We show that effective node-level motif detection requires GNNs to be flexible to multiple frequency bands, which help us understand the impact of heterophily on graph-level tasks, and assist us to design new models for graph classification.

Implications for Heterophilic Graph Classification Our theoretical analysis reveals why heterophily impacts graph-level tasks differently than node-level tasks. In node classification, heterophily typically manifests as a spatial-consistency graph property that can be partially addressed through HFD dynamics. However, in motif-based graph classification, heterophily patterns are *task-dependent* and spatially localized, creating three distinct challenges as mentioned in 2.1.

In a fine-grain clarification, we further divide the heterophily in motif-detection scenario into two genres: **Motif-agnostic heterophily** emerges when the background graph exhibits connectivity patterns independent of motif detection requirements, and **Motif-specific heterophily** arises because discriminative signals often require heterophilic patterns at motif boundaries, regardless of the background graph's structure. **Heterophily interference** occurs when mismatched patterns between motifs and backgrounds (*e.g.,* heterophilic motifs embedded in homophilic backgrounds) create conflicting optimization objectives for frequency-dominated approaches. This analysis directly motivates our experimental design in Section 4, where we systematically evaluate all four combinations of motif and background heterophily patterns to validate that mixed-frequency architectures outperform frequency-dominated approaches across diverse heterophily configurations.

4 Experimental Validation

In this section, we will verify our claims on both synthetic and real-world datasets to support the theoretical analysis. For both synthetic and real-world experiments, we use three different GNN models, standard GCN [11], gradient flow GCN (GCN$_{gf}$) [12] and Adaptive Channel Mixing GNN (ACM-GNN) [24], where standard GCN and GCN$_{gf}$ will provably lead to the frequency-dominated dynamic which enhances either low-frequency or high-frequency signals, while ACM-GNN are designed to adaptively combine multiple frequency filters. We make the parameters of these three models in the similar level to ensure them comparable. The train, validation and test sets are split by 80%, 10% and 10% in both synthetic and real-world experiments. The optimizer is Adam and the learning rate is set to 0.01 and 0.001 for synthetic and real-world experiments, respectively.

4.1 Synthetic Experiment

We conduct experiments on four synthetic dataset variants, each representing a different combination of backbone and motif connectivity patterns: homophilic-homophilic, homophilic-heterophilic, heterophilic-homophilic, and heterophilic-heterophilic (see Fig. 1(A) for demonstration and Appendix B for details). For each dataset variant, we train all three GNN models and evaluate the best-performing model (selected via validation) on the corresponding test set.

Table 1: Results on the pK_a dataset and Dirichlet Energy analysis.

Performance		Shrink Ratio of Normalized Dirichlet Energy			
Model	MSE ↓	Boundary ↑	R_2-NH	R-CH=O	R-C(=NH)NH2
GCN	3.08 ± 0.31	0.14 ± 0.01	0.13 ± 0.01	0.17 ± 0.02	0.16 ± 0.03
GCN_{gf}	3.24 ± 0.40	0.13 ± 0.01	0.19 ± 0.01	0.12 ± 0.01	0.11 ± 0.02
ACM-GNN	$\mathbf{2.32 \pm 0.38}$	1.77 ± 0.17	0.24 ± 0.08	0.14 ± 0.01	0.18 ± 0.04

We employ the empirical Maximum Mean Discrepancy (MMD) to quantify how effectively each GNN learns to distinguish between graph embeddings of different classes:

$$\widehat{\mathrm{MMD}}^2_\kappa(\{\mathbf{h}_i\},\{\mathbf{g}_j\}) = \frac{1}{p^2}\sum_{i,i'=1}^{p}\kappa(\mathbf{h}_i,\mathbf{h}_{i'}) + \frac{1}{q^2}\sum_{j,j'=1}^{q}\kappa(\mathbf{g}_j,\mathbf{g}_{j'}) - \frac{2}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\kappa(\mathbf{h}_i,\mathbf{g}_j) \quad (3)$$

where $\{\mathbf{h}_i\}_{i=1}^p$ and $\{\mathbf{g}_j\}_{j=1}^q \in \mathbb{R}^d$ represent p and q final graph embeddings randomly sampled from graphs with and without motifs in the test set, respectively, and κ denotes the RBF kernel function.

Figure 1(B) presents the MMD scores across different experimental conditions. Higher MMD scores indicate superior discriminative capability between graphs containing motifs versus those without, while lower scores suggest diminished separation ability. Across all four scenarios, ACM-GNN consistently achieves higher MMD scores compared to the baseline models, while the two frequency-dominated approaches exhibit similar performance levels. These results validate our theoretical claims regarding the limitations of frequency-dominated GNNs for motif detection tasks.

4.2 Real-world Experiment

To practically verify our claim on real-world graph-level tasks, we evaluated the baseline models on a newly collected dataset (see Appendix C for details). The dataset comprises 6,714 molecules with their corresponding pK_a (acid-base dissociation constant) values. The pK_a value quantifies the acidity or basicity of a molecule and is strongly influenced by specific functional groups, which correspond to distinctive motifs in the molecular graph structure. This constitutes a graph-level regression task where molecules serve as input and the objective is to predict the pK_a value of them.

Table 1 reports the experimental results. The mean squared error (MSE) is evaluated on the test set and averaged over five independent runs. Consistent with the synthetic data experiments, ACM-GNN achieves the best performance among the three GNN models. We further compute the shrink ratio of the normalized Dirichlet energy, $\frac{\mathcal{E}^{Dir}(\mathbf{F}(t))/\|\mathbf{F}(t)\|^2}{\mathcal{E}^{Dir}(\mathbf{F}(0))/\|\mathbf{F}(0)\|^2}$ where t denotes the evolution time of the dynamical system (*i.e.,* the number of layers). Since Dirichlet energy reflects differences across edges, we focus on two categories: boundary edges (those connecting motifs to the backbone graph) and intra-motif edges (within the three functional groups). The results show that GCN and GCN_{gf} yield low shrink ratios for both boundary and intra-motif edges, indicating global over-smoothing consistent with LFD dynamics. In contrast, ACM-GNN produces high shrink ratios on boundary edges and low shrink ratios within motifs, effectively sharpening boundaries while smoothing internal embeddings. This behavior explains the observed performance gap across models and supports our hypothesis on heterophily patterns in motif detection (Sec. 3.1).

5 Limitation, Conclusion and Future Work

6 Conclusion

Our theoretical and empirical analysis demonstrates that effective motif detection demands a spatially adaptive dynamic, rejecting monolithic low- or high-frequency dominated (LFD/HFD) regimes. Successful models must resolve a fundamental tension: performing intra-motif smoothing to unify constituent nodes while simultaneously sharpening boundaries to distinguish the motif from the wider graph structure. Our energy-based framework formalizes why globally frequency-dominated dynamics are ill-suited for this, revealing that their asymptotic convergence actively destroys the spectral signatures required to identify local patterns.

6.1 Limitation

Our work, while providing a foundational energy-based perspective, has several limitations that offer clear directions for future inquiry.

- **Task Scope:** Our theoretical and empirical analysis is primarily focused on **motif-based classification**. We use this task as a representative template to illustrate our core theory but do not conduct a formal analysis of the other two graph labeling schemes previously proposed. These other tasks, which may rely on different structural information, could present unique dynamic requirements not fully captured by our current study.
- **Empirical Generality:** The experimental validation is confined to a **single real-world dataset**. This inherently limits the assessment of our framework's **generality and robustness**. The conclusions drawn may not readily transfer to graphs with substantially different topological properties, such as varying degree distributions, clustering coefficients, sizes, or dataset domains.
- **Asymptotic vs. Finite-Layer Dynamics:** Our energy-based framework provides a powerful lens for analyzing the **asymptotic convergence** of graph dynamics. However, practical GNNs are almost always shallow, operating in a **finite-layer** regime (e.g., 2–8 layers). A potential gap exists between the properties predicted by our long-term asymptotic model and the actual transient behavior of these shallow architectures, which is ultimately responsible for their performance.
- **Architectural Scope:** The presented analysis implicitly centers on the dynamics characteristic of standard **Message-Passing Neural Networks (MPNNs)**. The extent to which our energy-based perspective and taxonomy of dynamics apply to other, increasingly popular architectures—particularly **Graph Transformers** which employ different mechanisms like global attention—has not been investigated.

7 Future Work

This energy-based perspective opens several avenues for future research. First, a deeper characterization of motif-specific dynamics promises to guide the principled design of specialized GNN architectures that excel at local structural pattern recognition. Second, extending our dynamic-based taxonomy to a broader class of graph-level tasks—and to other architectures like Graph Transformers—could provide a unified theory explaining performance variations across different models and problem settings. Such an analysis would clarify the nuanced role of heterophily at both the node and graph levels, paving the way for more robust and versatile graph learning models.

Acknowledgement

This study was funded by the European Union (DTRIP4H, No. 101188432) and iCANDOC Precision Cancer Medicine (PCM) pilot program from the Research Council of Finland.

References

[1] Pietro Bongini, Monica Bianchini, and Franco Scarselli. Molecular generative graph neural networks for drug discovery. *Neurocomputing*, 450:242–252, 2021.

[2] Stephen D. Boyles, Nicholas E. Lownes, and Avinash Unnikrishnan. Transportation Network Analysis, Volume I: Static and Dynamic Traffic Assignment, January 2025. arXiv:2502.05182 [math].

[3] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

[4] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, March 2009. Publisher: Nature Publishing Group.

[5] Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations*, 2019.

[6] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004.

[7] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, 1991.

[8] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, February 2010.

[9] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[10] Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy BR Hayter, Richard Vickers, Charles Roberts, Jian Tang, et al. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6):bbab159, 2021.

[11] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[12] Francesco Di Giovanni, James Rowbottom, Benjamin Paul Chamberlain, Thomas Markovich, and Michael M. Bronstein. Understanding convolution on graphs via energies. *Transactions on Machine Learning Research*, 2023.

[13] Zheng Gong, Guifeng Wang, Ying Sun, Qi Liu, Yuting Ning, Hui Xiong, and Jingyu Peng. Beyond homophily: robust graph anomaly detection via neural sparsification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2104–2113, 2023.

[14] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[15] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[16] Chenqing Hua, Sitao Luan, Minkai Xu, Zhitao Ying, Jie Fu, Stefano Ermon, and Doina Precup. Mudiff: Unified diffusion for complete molecule generation. In *Learning on Graphs Conference*, pages 33–1. PMLR, 2024.

[17] Chenqing Hua, Bozitao Zhong, Sitao Luan, Liang Hong, Guy Wolf, Doina Precup, and Shuangjia Zheng. Reactzyme: A benchmark for enzyme-reaction prediction. *Advances in Neural Information Processing Systems*, 37:26415–26442, 2024.

[18] Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, January 2005. Publisher: American Chemical Society.

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

[20] István A. Kovács, Katja Luck, Kerstin Spirohn, Yang Wang, Carl Pollis, Sadie Schlabach, Wenting Bian, Dae-Kyum Kim, Nishka Kishore, Tong Hao, Michael A. Calderwood, Marc Vidal, and Albert-László Barabási. Network-based prediction of protein interactions. *Nature Communications*, 10(1):1240, March 2019. Publisher: Nature Publishing Group.

[21] Qincheng Lu, Jiaqi Zhu, Sitao Luan, and Xiao-Wen Chang. Flexible diffusion scopes with parameterized laplacian for heterophilic graph learning. In *The Third Learning on Graphs Conference*, 2024.

[22] Sitao Luan, Chenqing Hua, Qincheng Lu, Liheng Ma, Lirong Wu, Xinyu Wang, Minkai Xu, Xiao-Wen Chang, Doina Precup, Rex Ying, et al. The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges. *arXiv preprint arXiv:2407.09618*, 2024.

[23] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Is heterophily a real nightmare for graph neural networks to do node classification? *arXiv preprint arXiv:2109.05641*, 2021.

[24] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Revisiting heterophily for graph neural networks. *Advances in neural information processing systems*, 35:1362–1375, 2022.

[25] Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. *Advances in Neural Information Processing Systems*, 36:28748–28760, 2023.

[26] Sitao Luan, Mingde Zhao, Xiao-Wen Chang, and Doina Precup. Break the ceiling: Stronger multi-scale deep graph convolutional networks. *Advances in neural information processing systems*, 32, 2019.

[27] Guoyong Mao, Runzhan Liu, and Ning Zhang. Study of electronegativity from a network perspective. *Scientific Reports*, 15(1):7154, February 2025. Publisher: Nature Publishing Group.

[28] Aurélien Mazurie, Danail Bonchev, Benno Schwikowski, and Gregory A. Buck. Evolution of metabolic network organization. *BMC Systems Biology*, 4(1):59, May 2010.

[29] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[30] Mark EJ Newman. Community detection and graph partitioning. *EPL (Europhysics Letters)*, 103(2):28003, 2013.

[31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[32] Md Abdur Rahaman, Zening Fu, Armin Iraji, and Vince Calhoun. A Deep Biclustering Framework for Brain Network Analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5075–5085, Seattle, WA, USA, June 2024. IEEE.

[33] Charlotte Ramon and Jörg Stelling. Functional comparison of metabolic networks across species. *Nature Communications*, 14(1):1699, March 2023. Publisher: Nature Publishing Group.

[34] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816, 2009.

[35] Christian Tantardini and Artem R. Oganov. Thermochemical electronegativities of the elements. *Nature Communications*, 12(1):2087, April 2021. Publisher: Nature Publishing Group.

[36] Hannu Toivonen, Ashwin Srinivasan, Ross D King, Stefan Kramer, and Christoph Helma. Statistical evaluation of the predictive toxicology challenge 2000–2001. *Bioinformatics*, 19(10):1183–1193, 2003.

[37] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[38] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413, February 2020.

[39] Mingjian Wen, Evan Walter Clark Spotte-Smith, Samuel M. Blau, Matthew J. McDermott, Aditi S. Krishnapriyan, and Kristin A. Persson. Chemical reaction networks and opportunities for machine learning. *Nature Computational Science*, 3(1):12–24, January 2023. Publisher: Nature Publishing Group.

[40] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning, 2018.

[41] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: Algorithms, applications and open challenges. In *International Conference on Computational Social Networks*, pages 79–91. Springer, 2018.

[42] Yilun Zheng, Xiang Li, Sitao Luan, Xiaojiang Peng, and Lihui Chen. Let your features tell the differences: Understanding graph convolution by feature splitting. In *The Thirteenth International Conference on Learning Representations*, 2025.

[43] Yilun Zheng, Sitao Luan, and Lihui Chen. What is missing for graph homophily? disentangling graph homophily for graph neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[44] Yilun Zheng, Zhuofan Zhang, Ziming Wang, Xiang Li, Sitao Luan, Xiaojiang Peng, and Lihui Chen. Rethinking structure learning for graph neural networks. *arXiv preprint arXiv:2411.07672*, 2024.

[45] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[46] Dongxiao Zhu and Zhaohui S. Qin. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics*, 6(1):8, January 2005.

[47] Jiong Zhu, Yujun Yan, Mark Heimann, Lingxiao Zhao, Leman Akoglu, and Danai Koutra. Heterophily and graph neural networks: Past, present and future. *IEEE Data Engineering Bulletin*, 2023.

A Real-world Examples for Graph-level Task Taxonomy

A.1 Examples for Aggregated Node Feature Based Labeling

- **Molecular property prediction**: Classifying drug molecules based on average atomic properties (*e.g.,* if the average electronegativity of atoms exceeds a threshold, classify as "polar" vs "non-polar") [27, 35]
- **Social network analysis**: Categorizing online communities based on aggregated user demographics (*e.g.,* if $> 70\%$ of users are in a certain age group, label the network as "young adult community") [37, 31]
- **Brain network analysis**: Classifying brain connectivity networks based on average activation levels across brain regions (*e.g.,* networks with high average activity labeled as "hyperactive state") [4, 32]
- **Protein interaction networks**: Classifying protein complexes based on the proportion of proteins belonging to specific functional categories [20, 16]

A.2 Substructure labeling examples

- **Drug discovery**: Classifying molecules based on the presence of specific pharmacophores or toxic substructures (*e.g.,* presence of benzene rings, specific functional groups) [1, 10]
- **Social network analysis**: Detecting communities based on local clustering patterns – networks with many tightly-knit triangular relationships vs. those with more heterophilic connections [34, 8]
- **Transportation networks**: Classifying road networks based on the presence of specific traffic patterns like roundabouts, highway interchanges, or bottleneck structures [2]
- **Chemical reaction networks**: Categorizing reaction pathways based on the presence of specific reaction motifs or catalytic cycles [17, 39]

A.3 Global labeling examples

- **Molecular classification**: Distinguishing between different molecular families based on overall structural properties like molecular diameter, overall connectivity, or graph density [40]
- **Social network analysis**: Classifying entire social networks based on global properties like average path length (small-world vs. random networks) or overall network density [15, 38]
- **Infrastructure networks**: Classifying power grids or communication networks based on their overall robustness, measured by global connectivity metrics [13]

B The Synthetic Dataset

Our synthetic dataset consists of graphs composed of a larger background graph (backbone) with a smaller substructure (motif) embedded within it. We generate backbone graphs and motifs independently, with each type designed to exhibit either homophilic or heterophilic node connectivity patterns.

Dataset Composition. We generate 1,000 backbone graphs for each connectivity type (homophilic and heterophilic) and create 5 distinct motif types, also categorized by connectivity pattern. Each backbone graph is paired with one motif to form a complete graph, resulting in four possible combinations: (homophilic backbone, homophilic motif), (homophilic backbone, heterophilic motif), (heterophilic backbone, homophilic motif), and (heterophilic backbone, heterophilic motif). With 1,000 backbone graphs and 5 motif variants per combination, each of the four combinations contains 5,000 graphs, yielding a total dataset of 20,000 graphs.

Graph Generation Process. The synthesis procedure consists of two stages: (1) structural generation of unlabeled graph skeletons, and (2) node feature assignment. The skeleton is first constructed by adding extra edges to a random tree, where the number of added edges is preset to half of the maximum possible edges for the given number of nodes. For homophilic graphs, node labels are assigned using the Clauset–Newman–Moore greedy modularity maximization algorithm [6], which

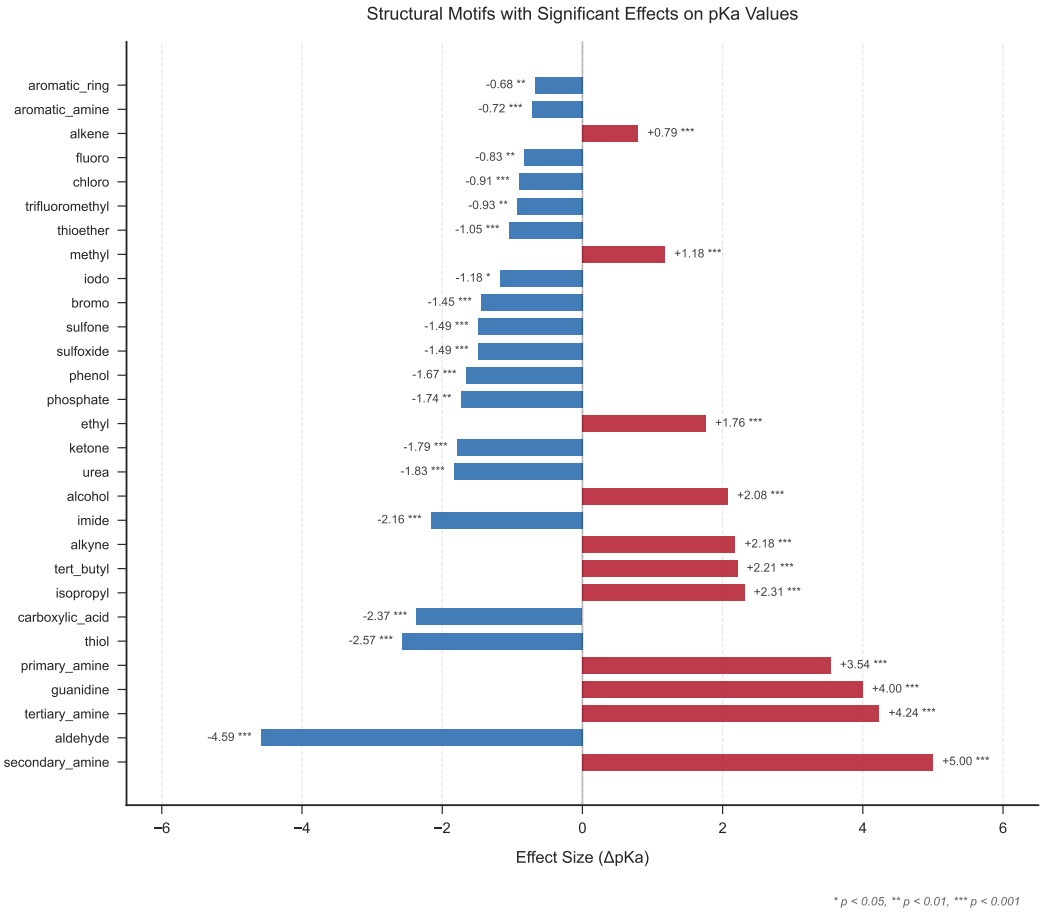

Figure 2: The influence of the functional groups on the pK_a values. The y-axis refers to different functional groups, and the x-axis refers to the change of the pK_a value if this specific functional group appears in the molecule.

encourages intra-community similarity. For heterophilic graphs, node labels are instead assigned at random to promote dissimilarity among connected nodes. The initial node features $\mathbf{f}_0$ are generated using a fixed embedding layer in PyTorch with small perturbations (gaussian noise with $\mu = 0$ and $\sigma = 0.05$), ensuring that nodes sharing the same label have similar features.

Graph Constraints. Backbone graphs contain 20–50 nodes, while motifs contain 5–7 nodes. And the edges are randomly sampled from minimum ($|V| - 1$) to maximum ($\frac{|V| \times (|V|-1)}{4}$). The maximum number of edges is set to ensure reasonable connectivity without creating overly dense graph.

C The pK_a Dataset

The dataset is collected from the IUPAC Dissociation Constants GitHub repository [2], which serves as a publicly accessible, continuously updated resource containing high-confidence pK_a data. This dataset has been meticulously digitized and curated from authoritative reference works published by the International Union of Pure and Applied Chemistry (IUPAC), ensuring data quality and reliability for computational chemistry applications.

From the total 24,290 records available in the repository, we select 6,714 unique molecules with their corresponding pK_aH_1 values, which represent the equilibrium constant for the loss of the first proton from each molecule. This subset focuses on molecules with well-defined ionization properties,

[2] `https://github.com/IUPAC/Dissociation-Constants`

Table 2: Statistical properties of the pK_a value in the dataset.

Average	Standard Variance	Maximum	Minimum
4.063	4.259	14.110	-17.632

making it suitable for studying the relationship between molecular structure and chemical reactivity. The pK_a values in our dataset span a wide range of chemical environments, encompassing various functional groups and molecular frameworks encountered in pharmaceutical and chemical research. In Fig. 2 we list the influence of functional groups on the molecule's pK_a.

For our task, we convert each molecule into a graph representation where atoms serve as nodes and chemical bonds as edges. Node features include atomic properties, *e.g.,* atomic number, formal charge, and hybridization state, while edge features capture bond types and stereochemistry information.

Table 2 presents statistical properties of the pK_a values in our curated dataset, including the distribution range, mean, median, and standard deviation, providing insight into the chemical diversity and complexity of the molecular structures under investigation.

D Equation Derivation

D.1 Eigenvectors Decomposition

Since $\mathbf{W} \in \mathbb{R}^{n\times n}$ is symmetric, we write it as $\mathbf{W} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{\Lambda}$ is a digonal matrix with all the eigenvalues $\boldsymbol{\mu} \in \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{n-1}\}$. Suppose $\boldsymbol{\mu}_k$ is the smallest positive eigenvalue, and $\boldsymbol{\mu}_0 \le \boldsymbol{\mu}_1 \le \cdots \le \boldsymbol{\mu}_k \le \cdots \le \boldsymbol{\mu}_{n-1}$, we have:

$$
\begin{aligned}
\mathbf{\Lambda} &= \begin{bmatrix} \boldsymbol{\mu}_0 & 0 & \ldots & 0 \\ 0 & \boldsymbol{\mu}_1 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & \boldsymbol{\mu}_{n-1} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{\mu}_0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \boldsymbol{\mu}_1 & \ldots & \ldots & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \boldsymbol{\mu}_{k-1} & \ldots & 0 \\ \vdots & \vdots & \ldots & \ldots & \ddots & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & \boldsymbol{\mu}_k & \ldots & 0 \\ \vdots & \vdots & \ldots & \ldots & \ddots & 0 \\ 0 & 0 & \ldots & \ldots & \ldots & \boldsymbol{\mu}_{n-1} \end{bmatrix} \\
&= \mathbf{\Lambda}_+ - \mathbf{\Lambda}_-
\end{aligned}
$$

where $\mathbf{\Lambda}_+$ is a diagonal matrix with all the positive eigenvalues (from $\boldsymbol{\mu}_k$ to $\boldsymbol{\mu}_{n-1}$) and $\mathbf{\Lambda}_-$ is a diagonal matrix with absolute values of all the negative eigenvalues (from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_{k-1}$). We note $\sqrt{\mathbf{\Lambda}_+}$ and $\sqrt{\mathbf{\Lambda}_-}$ as the element-wise square root of the two matrices $\mathbf{\Lambda}_+$ and $\mathbf{\Lambda}_-$ respectively. Furthermore, we define $\mathbf{\Theta}_+ = \sqrt{\mathbf{\Lambda}_+}\mathbf{Q}^\top$ and $\mathbf{\Theta}_- = \sqrt{\mathbf{\Lambda}_-}\mathbf{Q}^\top$. Then we can rewrite the weight matrix $\mathbf{W}$ as:

$$
\begin{aligned}
\mathbf{W} &= \mathbf{Q}(\mathbf{\Lambda}_+ - \mathbf{\Lambda}_-)\mathbf{Q}^\top \\
&= \mathbf{Q}\mathbf{\Lambda}_+\mathbf{Q}^\top - \mathbf{Q}\mathbf{\Lambda}_-\mathbf{Q}^\top \\
&= \mathbf{Q}(\sqrt{\mathbf{\Lambda}_+})(\sqrt{\mathbf{\Lambda}_+})\mathbf{Q}^\top - \mathbf{Q}(\sqrt{\mathbf{\Lambda}_-})(\sqrt{\mathbf{\Lambda}_-})\mathbf{Q}^\top \\
&= (\mathbf{Q}\sqrt{\mathbf{\Lambda}_+})(\sqrt{\mathbf{\Lambda}_+}\mathbf{Q}^\top) - (\mathbf{Q}\sqrt{\mathbf{\Lambda}_-})(\sqrt{\mathbf{\Lambda}_-}\mathbf{Q}^\top) \\
&= \mathbf{\Theta}_+^\top\mathbf{\Theta}_+ - \mathbf{\Theta}_-^\top\mathbf{\Theta}_-
\end{aligned}
$$

E Proofs

E.1 Proof of Proposition 1

Proof. We prove both directions of the equivalence.

(i) ⇒ (ii) Assume graph G contains a subgraph isomorphic to motif M. That is, there exists a subgraph $H = (V_H, E_H)$ where $V_H \subseteq V$, $E_H \subseteq E$, and $H \cong M$.

Since H is isomorphic to M, by definition, every node $i \in V_H$ is part of a subgraph in G that is isomorphic to M (namely, H itself). Therefore, by Definition 2, for every node $i \in V_H$, we have $y_i = 1$.

Since $V_H \neq \emptyset$ (as M is a non-empty motif), there exists at least one node $i \in V$ such that $y_i = 1$.

(ii) ⇒ (i) Assume there exists at least one node $i \in V$ such that $y_i = 1$.

By Definition 2, $y_i = 1$ means that node i is part of some subgraph in G that is isomorphic to M. Let H denote this subgraph. Then $H \subseteq G$ and $H \cong M$.

Therefore, graph G contains a subgraph isomorphic to motif M.

Conclusion Thus, the equivalence (i) ⇔ (ii) is established. This proves that the objective of distinguishing nodes in a substructure (node-level motif detection) is equivalent to detecting the existence of the substructure (graph-level motif detection) in the sense that:

$$\exists \text{ subgraph } H \subseteq G : H \cong M \quad \Leftrightarrow \quad \exists i \in V : y_i = 1$$

□

E.2 Proof of Lemma 2

Proof. To prove this, we must show two things: (1) a perfect node-level classifier contains sufficient information to construct a perfect graph-level classifier, and (2) a perfect graph-level classifier does not contain sufficient information to construct a perfect node-level classifier.

Let $G = (V, E)$ be a graph and M be a target motif. The node-level task requires learning a function $f_{node} : \mathcal{G} \to \{0, 1\}^{|V|}$ that predicts the label vector $\mathbf{y} = (y_1, \ldots, y_{|V|})$, where $y_i = 1$ if node i is part of a subgraph isomorphic to M. The graph-level task requires learning a function $f_G : G \to \{0, 1\}$ that predicts a single label $y_G = 1$ if G contains any subgraph isomorphic to M.

A node-level solution implies a graph-level solution. Given a perfect node-level classifier f_{node}, we can construct a perfect graph-level classifier, f_G^*, using a simple transformation $T : \{0, 1\}^{|V|} \to \{0, 1\}$:

$$f_G^*(G) = T(f_{node}(G)) = \max(f_{node}(G)).$$

By definition, a graph G contains a motif M if and only if there is at least one node $i \in V$ that is part of a subgraph isomorphic to M. This is equivalent to the condition that at least one entry in the true node-level label vector $\mathbf{y}$ is 1. The function f_G^* correctly implements this, as $\max(\mathbf{y}) = 1$ if and only if $\sum y_i > 0$. Thus, the information provided by f_{node} is sufficient to solve the graph-level task.

A graph-level solution does not imply a node-level solution. We prove this by counterexample. Let the motif M be a triangle (K_3) and consider two graphs, G_1 and G_2, on the vertex set $V = \{v_1, v_2, v_3, v_4\}$.

- Let G_1 be a graph consisting of a single triangle on nodes $\{v_1, v_2, v_3\}$ and an isolated node v_4.
- Let G_2 be a complete graph (K_4) on all four nodes.

A perfect graph-level classifier f_G will produce the same output for both, as both graphs contain at least one triangle:

$$f_G(G_1) = 1 \quad \text{and} \quad f_G(G_2) = 1.$$

However, the ground-truth node-level label vectors are different:

- For G_1, the node-level vector is $\mathbf{y}_1 = (1, 1, 1, 0)$.
- For G_2, the node-level vector is $\mathbf{y}_2 = (1, 1, 1, 1)$, since all nodes in a K_4 are part of at least one triangle.

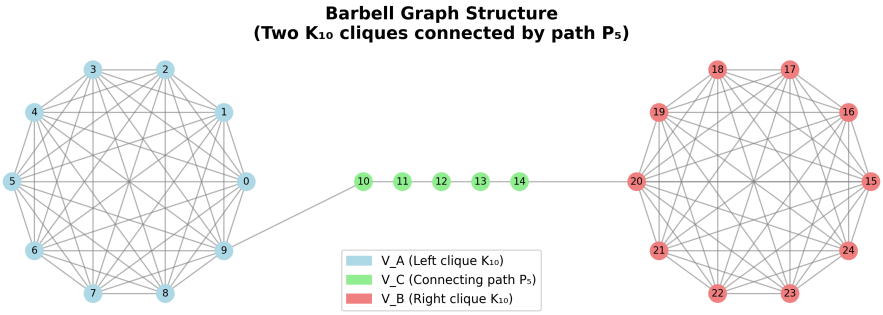

Figure 3: A barbell-shape graph consists of two complete graph K_{10} and a connection path line graph P_5.

Any transformation attempting to construct a node-level classifier from the graph-level output would need to map the input value '1' to two different outputs, $\mathbf{y}_1$ and $\mathbf{y}_2$. This is impossible for a function. Therefore, the information provided by a graph-level classifier is fundamentally insufficient to solve the node-level task, as it cannot distinguish between different node configurations that satisfy the same graph-level property. Since a node-level solution implies a graph-level solution but the reverse is not true, the node-level task is strictly more general. □

E.3 Proof of Theorem 1

Proof. Let $\mathcal{L}_{motif}$ denote the motif detection loss function, and let $\mathbf{f}_M^*$ denote the optimal node feature representation for motif detection. We analyze how the energy functional $\mathcal{E}_\theta(\mathbf{F})$ from Eq. 2 conflicts with motif detection objectives under LFD and HFD regimes.

LFD Suboptimality According to the framework in Section 2, LFD dynamics minimize the normalized Dirichlet energy $\mathcal{E}^{Dir}(\mathbf{F}(t))/\|\mathbf{F}(t)\|^2$, driving the system toward the global minimum where $\mathbf{F}(t) \to c\boldsymbol{\phi}_0\mathbf{1}^T$ (c is the proportional coefficient).

In the gradient flow formulation $\dot{\mathbf{F}}(t) = -\nabla\mathcal{E}_\theta(\mathbf{F}(t))$, the pairwise interaction term $-\sum_{i,j} A_{ij}\langle\mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle$ encourages neighboring nodes to have aligned representations (when $\mathbf{W}$ has positive eigenvalues). This global alignment objective is fundamentally at odds with motif detection.

For motif detection, the optimal loss $\mathcal{L}_{motif}$ requires representations to distinguish between motif-participating nodes V_M and non-participating nodes $V \setminus V_M$. However, LFD dynamics optimize for global consensus, producing identical representations $\mathbf{f}_i \approx \mathbf{f}_j$ for all $i, j \in V$ (*e.g.,* the node on the boundary like node 9 and node 10 in the Fig. 3). This makes the model to have blurred edges between motifs and the background graph and eliminates the discriminative information necessary for motif detection, resulting in $\mathcal{L}_{motif}(c\boldsymbol{\phi}_0\mathbf{1}^T) > \mathcal{L}_{motif}(\mathbf{f}_M^*)$.

HFD Suboptimality From the theorem in Section 2, HFD dynamics occur when $|\boldsymbol{\mu}_0|(\boldsymbol{\lambda}_{n-1} - 1) > \boldsymbol{\mu}_{d-1}$, leading to maximization of the normalized Dirichlet energy and convergence to $\mathbf{F}(t) \to c\boldsymbol{\phi}_{n-1}\mathbf{1}^T$.

In this regime, the energy functional's pairwise term $-\sum_{i,j} A_{ij}\langle\mathbf{f}_i, \mathbf{W}\mathbf{f}_j\rangle$ drives neighboring nodes to have maximally different representations, as the system seeks to maximize $\sum_{(i,j)\in E}\|\mathbf{f}_i - \mathbf{f}_j\|^2$.

This creates a fundamental conflict with motif detection objectives. Consider any connected motif M - nodes within the same motif instance are connected by edges and should receive similar labels (both should be classified as motif participants). However, HFD dynamics force these neighboring nodes to have anti-aligned representations: $\mathbf{f}_i \approx -\mathbf{f}_j$ for $(i, j) \in E$, *e.g.,* nodes in the complete graph K_{10} on the sides in Fig. 3.

The energy minimization process actively works against the motif detection objective, making it impossible for any linear classifier to consistently label connected nodes within the same motif instance. Thus $\mathcal{L}_{motif}(c\boldsymbol{\phi}_{n-1}\mathbf{1}^T) > \mathcal{L}_{motif}(\mathbf{f}_M^*)$.

Mixed-Frequency Requirement The energy framework reveals why motif detection is incompatible with frequency-dominated regimes. The gradient flow $\dot{\mathbf{F}}(t) = -\nabla\mathcal{E}_\theta(\mathbf{F}(t))$ drives the system toward eigenvector alignment, but motif detection requires a different objective function altogether.

Optimal motif detection requires minimizing $\mathcal{L}_{motif}$, which demands:

- **Inner smoothing**: Similar representations within motif instances, *i.e.,* local low-frequency behavior
- **Boundary sharpening**: Sharp boundaries between motif and non-motif regions, *i.e.,* local high-frequency behavior
- **Scale-appropriate sensitivity**: Frequency components matching the motif's characteristic size

This creates a mixed-frequency optimization problem that cannot be solved by the global energy minimization of $\mathcal{E}_\theta$. The optimal solution $\mathbf{f}_M^*$ requires spatially-varying frequency content: $\mathbf{f}_M^* = \sum_k \alpha_k \boldsymbol{\phi}_k \mathbf{v}_k$ where the coefficients α_k depend on the local structural context around each motif.

Since the energy functional in Eq. 2 enforces global spectral alignment (either LFD or HFD), it cannot accommodate the spatially-heterogeneous frequency requirements of motif detection. Therefore, effective motif detection requires architectures that can escape the LFD/HFD dichotomy imposed by the standard energy framework. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The full code/dataset will be published in the future complete version of this work.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.