Leap of FAITH from GNN-to-MLP: Fairness Aware Inference via DisTillation of GrapH Knowledge

Vipul Kumar Singh[1*], **Jyotismita Barman**[1*], **Sandeep Kumar**[1,2,3,4], **Tapan K. Gandhi**[1,2,3], **Jayadeva**[1,2]

[1]Department of Electrical Engineering, Indian Institute of Technology, Delhi, India
[2]Yardi School of Artificial Intelligence, Indian Institute of Technology, Delhi, India
[3]Bharti School of Telecommunication Technology and Management, Indian Institute of Technology, Delhi, India
[4]Biofin Capital, Princeton, New Jersey, United States of America
{Vipul.Kumar.Singh, Jyotismita.Barman, ksandeep, tgandhi, jayadeva}@ee.iitd.ac.in

Abstract

Graph Neural Networks (GNNs) are expressive architectures for learning from complex graph-structured data. However, their practical use is often limited by the high computational cost of neighborhood aggregation. Recent efforts have focused on knowledge distillation from GNNs to inference-efficient Multi-Layer Perceptrons (MLPs). However, most existing works treat this distillation as an embedding alignment problem, overlooking the need to replicate the topology-aware smoothing behavior that arises from message passing in GNNs. Moreover, existing methods are primarily performance driven, ignoring critical real-world requirements such as fairness. In this work, we make two key observations: *(1)* state-of-the-art distillation methods fail to capture the heterogeneous smoothness patterns of GNNs, limiting structural awareness in MLPs, and *(2)* they introduce significant individual and group fairness violations. We introduce FAITH, the first *fair and structurally aware GNN-to-MLP distillation framework with graph-free inference.* To improve structural awareness in MLPs, we propose a neighborhood-guided energy alignment objective that transfers not only node-level energy, but also the distribution of energies across local neighborhoods. To improve individual fairness, FAITH introduces a novel $\ell_{2,1}$-norm objective that preserves structured similarity in the learned representations. Additionally, we incorporate a counterfactual invariance objective that explicitly encourages the model to learn representations that are statistically independent of the sensitive attribute. We provide a comprehensive theoretical analysis of FAITH, interpreting it through a novel instantiation of the Information Bottleneck principle. Extensive experiments on 11 benchmark datasets show that FAITH achieves stronger structural awareness and delivers a better trade-off between utility and fairness than existing methods.

Code & Appendix — https://github.com/shashivipul/FAITH

Introduction

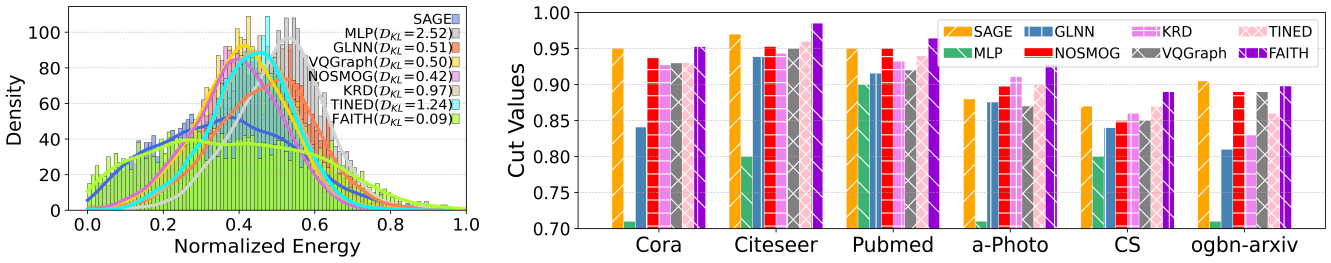
Graph Neural Networks (GNNs) have demonstrated remarkable potential in learning representations for graph-structured data (Kipf and Welling 2016; Hamilton, Ying, and Leskovec 2017). By leveraging connectivity guided representation learning, GNNs have found widespread application in domains such as drug discovery, epidemic modeling, molecular analysis, and recommender systems (Zhang et al. 2021; Reiser et al. 2022; Liu et al. 2024). Despite these successes, the reliance on neighborhood aggregation in GNNs presents significant challenges for large-scale and time-sensitive applications (Liu et al. 2022). In contrast, Multi-Layer Perceptrons (MLPs) remain the industry standard for deployment on edge devices and in practical pipelines due to their reduced computational cost and latency (Zhou et al. 2024). However, traditional MLPs are unable to leverage relational information within the data, significantly limiting their ability to model neighborhood interactions.

Recent studies have attempted to transfer knowledge from GNN teachers to inference-efficient MLPs via knowledge distillation (Liang et al. 2023; Zhang et al. 2022; Tian et al. 2022; Wu et al. 2023). These methods primarily align the representation distributions of the teacher and student models. Despite significant progress, existing methods overlook two critical challenges: *(1)* the inability to capture the heterogeneous smoothness patterns of teacher GNNs, which are crucial for expressive power over graph structures, and *(2)* the presence of biases in the student representations.

Challenge 1. GNNs with neighborhood aggregation inherently act as node signal denoisers, producing smooth representations (Ma et al. 2021). The graph Dirichlet Energy quantifies the smoothness of these node signals. However, this smoothness is not uniform; it varies spatially across the graph to balance expressivity. Results presented in Figure 1(a) highlight that existing distillation methods fail to replicate the heterogeneous energy distribution captured by the teacher GNNs. This limitation is further evidenced by lower Cut Value ($\mathcal{CV} \in [0, 1]$) scores (Zhang et al. 2022), as shown in Figure 1(b).

Challenge 2. GNNs have been shown to encode discriminatory biases in their representations (Dong et al. 2023; Singh et al. 2024). Therefore, blindly forcing the student to imitate teacher output may transfer and even amplify these biases in the student representations. The results in Figure 2 confirm the presence of (un)-fairness in both teacher GNN and the

*These authors contributed equally.



(a) **Energy distribution** (b) **Cut Values**

Figure 1: **[Challenge 1] (a)** Histogram and KL-divergence ($\mathcal{D}_{KL}$) values of node-level energy distributions for the teacher GNN and various distilled MLPs on the `Cora` dataset. Existing methods collapse the distribution, failing to capture the heterogeneous smoothness patterns of the teacher. **(b)** Cut Values (↑) comparison between the teacher GNN and distilled MLPs across multiple datasets, showing that existing methods fall short of matching the structural consistency achieved by the GNN.

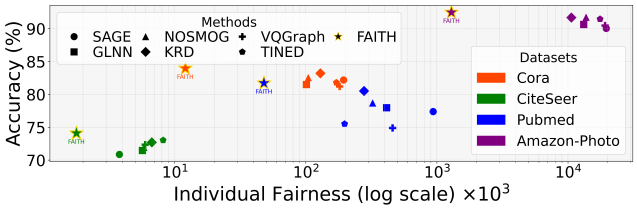


Figure 2: **[Challenge 2]** Comparison of utility and individual (un)-fairness between GNNs and distilled student models on different datasets. For each dataset, the optimal method lies closest to the top-left corner. Fairness values are in thousands.

Method	Energy Alignment	Neighbourhood Guidance	Graphless Inference	Fairness Objective
GLNN	✗	✗	✓	✗
NOSMOG	✗	✓	✗	✗
KRD	✗	✓	✓	✗
VQGRAPH	✗	✓	✓	✗
TINED	✓	✗	✓	✗
FAITH	✓	✓	✓	✓

Table 1: Characterization of existing GNN-to-MLP distillation methods. A ✓ indicates the presence of a desirable property, and ✗ indicates its absence.

distilled MLPs, highlighting the need for caution when transferring knowledge from unreliable teacher. Hence, we pose the research problem of acquiring MLPs with neighborhood-aware smoothing for graph-free inference. Additionally, we aim to immunize the distilled MLPs against bias propagation from an unreliable teacher.

Design Choice Analysis. We next outline the key design choices required to address the above research problem. These include: (1) Energy Alignment, (2) Neighborhood Guidance, and (3) Fairness-Promoting Objective. Table 1 summarizes the extent to which these design choices are adopted or neglected by existing methods. GLNN (Zhang et al. 2022) applies soft-label matching but ignores all key design choices. NOSMOG (Tian et al. 2022) appends DeepWalk features to incorporate topology, but depends on graph access during inference, limiting its use in graph-free settings. KRD (Wu et al. 2023) selects reliable nodes via perturbation-invariant entropy and aligns student embeddings with those of the teacher and its reliable neighbors. VQGraph (Yang et al. 2024) encodes substructures using a pretrained GNN-based tokenizer and codebook, but requires access to the teacher architecture and codebook pretraining. TINED (Zhou et al. 2025) aligns global Dirichlet energy but overlooks local smoothness, leading to poor structural alignment. This is reflected by its higher divergence from the energy distribution of the teacher model and lower Cut Values in Figure 1. Notably, none of these methods incorporate fairness-enhancing objectives. Our key contributions are as follows:

1. **Topology-Guided Energy Distillation.** We identify that existing GNN-to-MLP distillation methods fail to align the Dirichlet energy distributions of student MLPs with those of their teacher GNNs. To address this, we propose a neighborhood-guided energy alignment objective that transfers both node-level and neighborhood-level smoothness patterns to the student. This mitigates the challenge of performing heterogeneous smoothing for expressivity. As demonstrated in Figure 1, `FAITH` achieves the closest energy distribution to the teacher and the highest Cut Value scores.
2. **Fairness-Aware Distillation.** MLPs distilled from unreliable GNNs often exhibit individual and group (un)-fairness. We address this via a structured similarity-preserving $\ell_{2,1}$-norm regularizer to control Lipschitz behavior, and a counterfactual invariance objective to encourage group fairness. This prevents the transfer of bias to the MLPs during knowledge distillation from the teacher. As shown in Figure 2, `FAITH` significantly reduces bias in the representations while also improving predictive utility.
3. **Comprehensive Evaluation.** Experiments on 11 benchmark datasets show that `FAITH` outperforms state-of-the-art methods in balancing utility and fairness.

Background and Preliminaries

Notations. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be an undirected graph, where $\mathcal{V}$ is the node set, $\mathcal{E}$ the edge set, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the node feature matrix with $n = |\mathcal{V}|$ and feature dimension d. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ satisfies $\mathbf{A}_{uv} = 1$ if $(u, v) \in \mathcal{E}$,

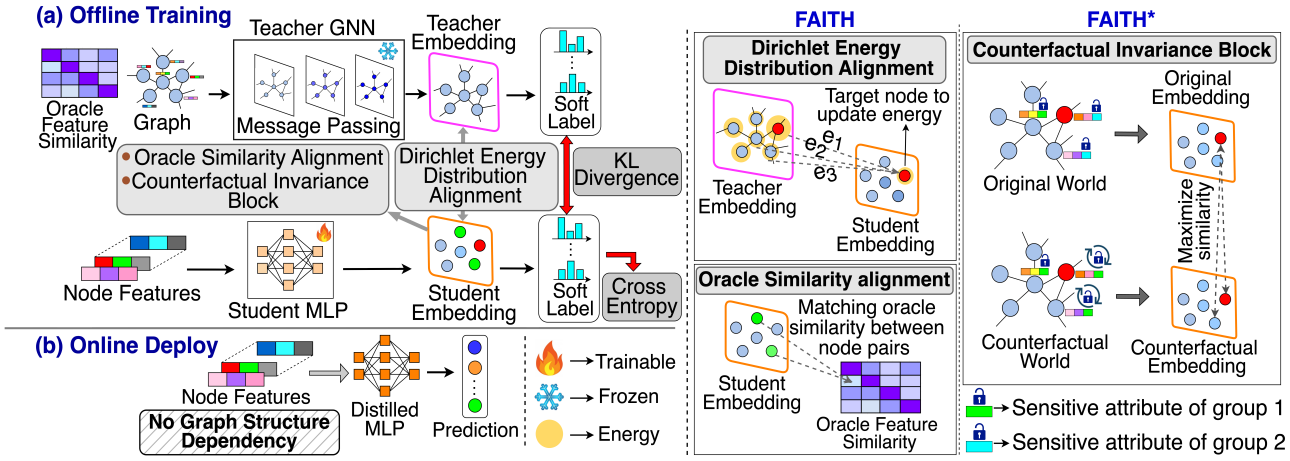


Figure 3: Overview of FAITH workflow: (a) The MLP is trained by distilling knowledge from a pre-trained GNN. (b) During inference, the MLP operates independently, utilizing only node features for predictions.

and 0 otherwise. For a node $v \in \mathcal{V}$, its 1-hop neighborhood is $\mathcal{N}(v) = \{u \in \mathcal{V} : (v, u) \in \mathcal{E}\}$. $\mathbf{D}$ denotes degree matrix.

Problem Setup. We consider semi-supervised node classification with label matrix $\mathbf{Y} \in \mathbb{R}^{n \times C}$, where C is the number of classes. The labeled node set $\mathcal{V}^L \subset \mathcal{V}$ has features $\mathbf{X}^L$ and labels $\mathbf{Y}^L$; the goal is to predict labels $\mathbf{Y}^U$ for unlabeled nodes $\mathcal{V}^U = \mathcal{V} \setminus \mathcal{V}^L$ using features $\mathbf{X}^U$.

Fair Representation Learning. Algorithmic fairness is typically categorized into three primary notions: (1) *individual fairness*, which requires that similar individuals receive similar predictions; (2) *group fairness*, which seeks to ensure equitable treatment across different sensitive groups; and (3) *counterfactual fairness*, which demands prediction invariance under counterfactual changes to sensitive attributes.

Definition 1 (Individual Fairness)**.** Let x_i and x_j be two distinct nodes in a graph. The model outputs $f(x_i)$ and $f(x_j)$ are said to be *individually fair* with respect to a similarity function $\mathcal{S}$ and a distance metric D if the condition holds:

$$D(f(x_i), f(x_j)) \leq \frac{\epsilon}{\mathcal{S}(x_i, x_j)} \quad \forall\, i \neq j \tag{1}$$

where ϵ is a parameter controlling fairness tolerance, and the similarity measure satisfies $0 \leq \mathcal{S}(x_i, x_j) \leq 1$.

Definition 2 (Statistical Parity)**.** Statistical parity (SP), also referred as demographic parity, evaluates whether predictions of a model are independent of a sensitive attribute that defines group membership. It is measured as the difference in the probability of assigning a positive label across two groups defined by this attribute:

$$\Delta\text{SP} = \mathbb{P}(\hat{y} = 1 \mid G = 0) - \mathbb{P}(\hat{y} = 1 \mid G = 1). \tag{2}$$

Let $\hat{y}$ be the predicted outcome and $G \in \{0, 1\}$ a binary sensitive attribute indicating group membership. Statistical parity holds when $\Delta\text{SP} = 0$, meaning both groups receive positive predictions at equal rates. This implies statistical independence between $\hat{Y}$ and G, promoting group fairness.

FAITH Methodology

This section introduces the FAITH (**F**airness **A**ware **I**nference via dis**T**illation of grap**H** knowledge) framework, which aligns Dirichlet energy distributions between teacher and student networks while simultaneously enhancing fairness in predictions. Figure 3 illustrates the key components of the framework. FAITH introduces: *(1)* a Dirichlet Energy Distribution (DED) alignment block to enhance the structural awareness of the student MLP, and *(2)* an Oracle Similarity Alignment (OSA) block to enforce individual fairness (IF). To further address group fairness, we incorporate a Counterfactual Invariance Block into FAITH, defining the extended framework, FAITH⋆.

Neighborhood-Guided Energy Distillation. GNNs perform topology-aware signal denoising through neighborhood aggregation (Ma et al. 2021). Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$, the graph Dirichlet energy is defined as:

$$E = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} e_i; \quad e_i = \frac{1}{2\sqrt{|\mathcal{N}(i)|d}} \sum_{j \in \mathcal{N}(i)} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2, \tag{3}$$

where e_i represents the node-level energy, which quantifies the local smoothness of node features over the graph.

As shown in Figure 1(a), GNNs learn a heterogeneous distribution of node-level energy to balance expressivity and oversmoothing, resulting in higher Cut Values that reflect stronger structural consistency. In contrast, MLPs lack topology-guided supervision and produce flattened energy distributions, failing to capture this heterogeneity. TINED (Zhou et al. 2025) introduces global Dirichlet energy alignment but ignores local smoothness variations, leading to weak structural alignment and lower Cut Values (Figure 1(b)).

An alternative is to penalize the sum of node-wise energy deviations. However, this approach treats nodes in isolation and overlooks the shared structure within overlapping neighborhoods, limiting its ability to preserve local consistency. Effective representation learning requires leveraging the energy patterns of the teacher not only at the node level but

also across its neighborhood, ensuring node embeddings are updated with localized structural information.

To distill topology-aware node-level energy from GNNs to MLPs, we propose leveraging the neighbors of each target node $i \in \mathcal{V}$ as multiple sources of structural supervision. Let $\mathbf{z}_i^{\text{GNN}}$ and $\mathbf{z}_i^{\text{MLP}}$ denote the embeddings of node i in the GNN and the student MLP, respectively. We define the neighborhood-guided energy alignment loss as

$$\mathcal{L}_{\text{DE}} = \mathbb{E}_{i\in\mathcal{V}} \ \mathbb{E}_{j\in\mathcal{N}(i)\cup\{i\}} \left(e_i^{\text{MLP}} - e_j^{\text{GNN}}\right)^2. \tag{4}$$

This encourages the student MLP not only to match the energy of each node with its teacher counterpart but also to capture the energy patterns present in its local neighborhood.

Oracle Similarity Alignment. Based on Definition 1, individual fairness requires that the similarity structure induced by the oracle similarity function $S(x_i, x_j)$ is preserved in the output space of the model. Specifically, for all $x_i, x_j \in \mathcal{X}$, the predictions must satisfy $D(f(x_i), f(x_j)) \leq \epsilon / S(x_i, x_j)$. Although this condition provides a formal fairness guarantee, enforcing it exactly across all instance pairs is computationally infeasible in practice. A widely adopted relaxation is to convert this constraint into a pairwise smoothness regularization on the learned representations. Existing methods commonly use ℓ_2-norm smoothness objectives to promote fairness by aligning similar pairs (Kang et al. 2020; Song et al. 2022). But they suffer from these key limitations: *(i)* sensitivity to noise and outliers, *(ii)* excessive smoothing that masks data heterogeneity, and *(iii)* poor handling of nonlinear structure. The piecewise structure of real-world data is better preserved by less aggressive ℓ_1-norm penalties (Wang et al. 2016). However, they treat each feature dimension independently and fail to capture interactions across dimensions, which are essential for accurately modeling fairness (Ghosh, Basu, and Meel 2022).

To address the limitations of uniform ℓ_2- and ℓ_1-based smoothing, we introduce a structured sparsity-inducing regularization framework based on the $\ell_{2,1}$-norm. It enables more effective exploitation of feature correlations while selectively enforcing similarity across fairness-relevant points. Specifically, let $\mathcal{T}$ denote the set of sample pairs (i, j) for which $S_{ij} > 0$. We define the incidence matrix $\Pi \in \mathbb{R}^{|\mathcal{T}|\times n}$, where each row corresponds to a pair $(i, j) \in \mathcal{T}$ as:

$$\Pi[e, k] = \delta_{k,i} - \delta_{k,j}, \tag{5}$$

where $\delta_{a,b} = 1$ if $a = b$, and 0 otherwise. For pair index e corresponding to (i, j). Given the embedding matrix $Z \in \mathbb{R}^{n\times d}$, we define our $\ell_{2,1}$-norm based regularizer as:

$$\mathcal{L}_{\text{IF}} = \|W\Pi Z\|_{2,1} = \sum_{(i,j)\in\mathcal{T}} S_{ij}\|z_i - z_j\|_2 \tag{6}$$

where $W \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$ is diagonal matrix with $W[e, e] = S_{ij}$.

Counterfactual Invariance for Group Fairness. While the graph smoothness prior enables GNNs to learn expressive representations, it may also amplify group-level biases when topology aligns with sensitive attributes (Dai and Wang 2021). By transferring this inductive bias to MLPs through knowledge distillation, such discriminatory biases can propagate into student representations. To promote group fairness, we encourage the learned representations to be invariant to sensitive attributes by aligning the embeddings of original and counterfactual samples. Prior work has shown that enforcing representation consistency across counterfactual pairs can improve demographic parity (Rosenblatt and Witter 2023). Rather than generating counterfactuals synthetically or randomly, we exploit the graph structure to construct topology-aware counterfactuals.

Let $x_i \in \mathcal{V}$ be a node with sensitive attribute $G_i \in \{0, 1\}$. We define a contextual neighborhood:

$$\mathbb{C}_i = \{x_j \in \mathcal{N}(i) \cup \{x_i\} \mid G_j = G_i\}, \tag{7}$$

which captures the local distribution $p(X \mid g_i)$ under the assumption that neighbors share similar features, as implied by graph smoothness priors (Rue and Held 2005).

For each $x_j \in \mathbb{C}_i$, we construct a counterfactual x_j^- by flipping its sensitive attribute $G_j \leftarrow 1 - G_j$, holding all other features fixed. Let $z_j^- = f_{\text{MLP}}(x_j^-)$ denote the embedding under the student model. We define the counterfactual alignment loss as:

$$\mathcal{L}_{\text{GF}} = \mathbb{E}_{x_i\in\mathcal{V}} \ \mathbb{E}_{x_j\in\mathbb{C}_i} \left[1 - \frac{z_i^\top z_j^-}{\|z_i\| \cdot \|z_j^-\|}\right], \tag{8}$$

which encourages the student to learn representations invariant to changes in group membership.

Model Optimization. The final training objective of FAITH combines the following loss terms: supervised cross-entropy loss, soft-label distillation loss, neighborhood-guided energy alignment loss, and oracle-based similarity alignment loss. The overall objective is given by

$$\begin{aligned}\mathcal{L}_{\text{FAITH}} = \alpha \sum_{i\in\mathcal{V}^L} \mathcal{L}_{\text{CE}}\left(\sigma\left(\mathbf{z}_i^{\text{MLP}}\right), \mathbf{y}_i\right) + \beta\mathcal{L}_{\text{DE}} + \gamma\mathcal{L}_{\text{IF}} \\ + (1-\alpha)\sum_{i\in\mathcal{V}} \mathcal{D}_{\text{KL}}\left(\sigma\left(\mathbf{z}_i^{\text{MLP}}/\tau\right) \ \| \ \sigma\left(\mathbf{z}_i^{\text{GNN}}/\tau\right)\right),\end{aligned} \tag{9}$$

where $\sigma(\cdot)$ denotes the softmax function and τ is the temperature parameter used for distillation. $\mathcal{L}_{\text{CE}}(\cdot)$ and $\mathcal{D}_{\text{KL}}(\cdot)$ are the cross-entropy and KL-divergence loss functions, respectively. To extend this framework for handling both individual and group fairness constraints, the optimization objective for FAITH* includes an additional group fairness regularization term:

$$\mathcal{L}_{\text{FAITH}^\star} = \mathcal{L}_{\text{FAITH}} + \gamma^\star\mathcal{L}_{\text{GF}}, \tag{10}$$

where $\alpha, \beta, \gamma, \gamma^\star$ are weights used to balance the individual objectives. Algorithmic details are provided in Appendix E.

Theoretical Analysis of FAITH

In this section, we interpret the FAITH learning objectives and expressiveness from an information-theoretic view.

Fairness-Aware Knowledge Distillation under the Information Bottleneck Principle. The Information Bottleneck (IB) framework (Tishby and Zaslavsky 2015) provides a principled way to learn representations that are both compact and predictive. Given input $\mathcal{G}$ and target Y, the goal is to learn

a latent representation $\mathbf{Z}$ that discards irrelevant input information $I(\mathbf{Z};\mathcal{G})$ while preserving task-relevant information $I(\mathbf{Z};Y)$. This trade-off is captured by the Lagrangian:

$$\mathbf{Z}^\star = \arg\min_{\mathbf{Z}} \; I(\mathbf{Z};\mathcal{G}) - \lambda_1 I(\mathbf{Z};Y), \quad \lambda_1 \geq 0. \tag{11}$$

MLP-based encoders lack structure information, limiting their ability to capture relational dependencies. In addition, directly optimizing the standard IB objective can result in biased representations. To address both challenges, we extend the classical IB framework to incorporate fairness and alignment with teacher GNNs. We introduce a novel IB-based objective called Fair Knowledge Distillation via Information Bottleneck (`FairKD-IB`).

Problem 1 (`FairKD-IB` Principle). Let $f_{\text{GNN}} : \mathcal{X} \times \mathcal{A} \to \mathbf{Z}_{\text{T}}$ be a pretrained GNN encoder and $f_{\text{MLP}} : \mathcal{X} \to \mathbf{Z}_{\text{S}}$ a student model. Let $\mathcal{S}$ denote an oracle similarity function over input pairs, D a distance metric over student representations, and G a sensitive attribute. The `FairKD-IB` objective seeks a student representation that solves:

$$\begin{aligned}\mathbf{Z}^\star = \arg\min_{\mathbf{Z}_{\text{S}}} & I(\mathbf{Z}_{\text{S}};\mathcal{G}) - \lambda_1 I(\mathbf{Z}_{\text{S}};Y) - \lambda_2 I(\mathbf{Z}_{\text{T}};\mathbf{Z}_{\text{S}}) \\ & - \lambda_3 I(D(\mathbf{Z}_{\text{S}});\mathcal{S}) + \lambda_4 I(\mathbf{Z}_{\text{S}};G),\end{aligned} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geq 0$ are Lagrangian multipliers.

Proof. Proofs in the paper are provided in Appendix A. □

Proposition 1 (`FAITH` Optimizes a Relaxation of `FairKD-IB`)**.** *The optimization objectives $\mathcal{L}_{FAITH}$ and $\mathcal{L}_{FAITH^\star}$ correspond to a tractable relaxation of the `FairKD-IB` formulation. The mappings of the surrogate terms are provided in Appendix A.*

Proposition 2 (Fairness of Downstream Classifiers)**.** *Let $\Theta : \mathbb{R}^h \to \mathbb{R}^C$ be a downstream classifier with Lipschitz constant $\mathbb{L}_\Theta$. Suppose:*

1. *The encoder satisfies $D(f_{MLP}(x_i), f_{MLP}(x_j)) \leq \epsilon / S(x_i, x_j)$ for all input pairs (x_i, x_j), where S is a similarity function and D is a distance metric;*
2. *The representation $z = f_{MLP}(x) \in \mathbb{R}^h$ is statistically independent of the sensitive attribute G.*

Then, the downstream classifier Θ satisfies:

(a) $D(\Theta(z_i), \Theta(z_j)) \leq \mathbb{L}_\Theta \cdot \epsilon / S(x_i, x_j)$;
(b) $\Theta(z) \perp G$.

Expressivity Analysis. Let $\mathcal{G}^{(\nu)} = (\mathbf{X}^\nu, \mathbf{A}^{(\nu)})$ denote the ego-graph centered at node ν, with label $Y^{(\nu)}$. Training a model to predict $Y^{(\nu)}$ from $\mathcal{G}^{(\nu)}$ via cross-entropy is equivalent to maximizing the mutual information $I(\mathcal{G}^{(\nu)}; Y^{(\nu)})$. Treating $\mathcal{G}^{(\nu)}$ as a joint distribution over features and structure, this decomposes as: $I(\mathcal{G}^{(\nu)}; Y^{(\nu)}) = I(\mathbf{A}^{(\nu)}; Y^{(\nu)}) + I(\mathbf{X}^{(\nu)}; Y^{(\nu)} \mid \mathbf{A}^{(\nu)})$, which is the effective objective optimized by GNNs. In contrast, standard MLPs only maximize $I(\mathbf{X}^{(\nu)}; Y^{(\nu)})$, and thus lack relational expressivity. Distillation methods attempt to bridge this gap. Typical GNN-to-MLP distillation objectives take the form $\mathcal{L} = I(\mathbf{X}^{(\nu)}; Y^{(\nu)}) + I(\mathbf{X}^{(\nu)}; \mathbf{Z}_{\text{T}}^{(\nu)} \mid \mathbf{A}^{(\nu)})$, where $\mathbf{Z}_{\text{T}}^{(\nu)}$ denotes the teacher embedding. Approaches such as NOSMOG further inject structural signals into MLPs by augmenting node features with positional encodings.

`FAITH` introduces an additional objective that maximizes the mutual information $I(\mathbf{X}^{(\nu)}; \Psi(\mathbf{A}^{(\nu)}, \mathbf{Z}_{\text{T}}^{(\nu)}))$, where $\Psi(\mathbf{A}^{(\nu)}, \mathbf{Z}_{\text{T}}^{(\nu)})$ denotes the localized energy pattern derived from the graph structure and teacher embeddings.

Proposition 3. *Let $\Psi(\mathbf{A}^{(\nu)}, \mathbf{Z}_T^{(\nu)})$ be a localized energy pattern derived from the graph structure $\mathbf{A}^{(\nu)}$ and teacher embeddings $\mathbf{Z}_T^{(\nu)}$. Then the following inequality holds:*

$$\begin{aligned} I(\mathbf{X}^{(\nu)}; \Psi(\mathbf{A}^{(\nu)}, \mathbf{Z}_T^{(\nu)})) &\leq I(\mathbf{X}^{(\nu)}; \mathbf{A}^{(\nu)}, \mathbf{Z}_T^{(\nu)}) \\ &= I(\mathbf{X}^{(\nu)}; \mathbf{A}^{(\nu)}) + I(\mathbf{X}^{(\nu)}; \mathbf{Z}_T^{(\nu)} \mid \mathbf{A}^{(\nu)}). \end{aligned} \tag{13}$$

Proposition 3 implies that the `FAITH` objective implicitly encourages the preservation of structural information, since it maximizes a lower bound on the joint mutual information between the input features, graph topology, and teacher embeddings.

To assess structural alignment, we adopt the cut value metric (Zhang et al. 2022), defined as $\mathcal{CV} = \frac{\text{tr}(\hat{\mathbf{Y}}^\top \mathbf{A} \hat{\mathbf{Y}})}{\text{tr}(\hat{\mathbf{Y}}^\top \mathbf{D} \hat{\mathbf{Y}})}$, where $\hat{\mathbf{Y}}$ denotes the model predictions. Higher values indicate better consistency with the graph structure. As shown in Figure 1(b), `FAITH` achieves the highest cut value, demonstrating its superior ability to preserve topological information.

Results

This section presents comprehensive experiments to evaluate the effectiveness of `FAITH` on real-world node classification datasets, guided by the following research questions:

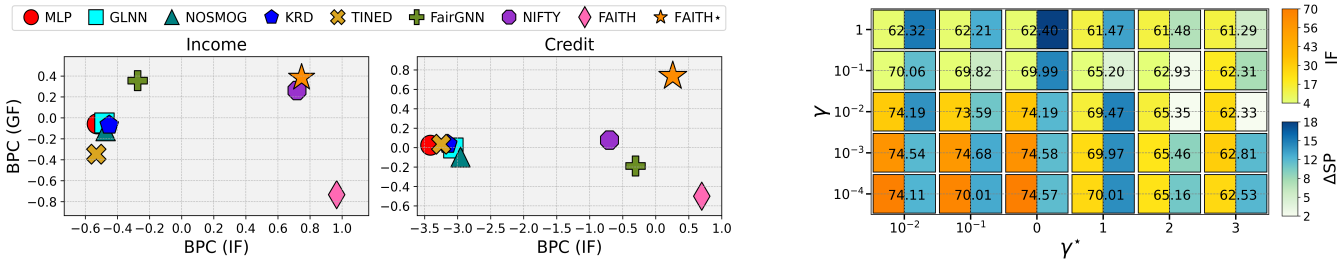
RQ1: *How well does `FAITH` balance utility and individual fairness across real world graph datasets in transductive and inductive settings?*
RQ2: *How does `FAITH`$\star$ maintain both group and individual fairness on datasets with sensitive attributes?*
RQ3: *How sensitive is the performance of `FAITH` to the choice of teacher GNN architecture?*
RQ4: *What are the effects of individual components and key hyperparameters in `FAITH`?*
RQ5: *How well does `FAITH` balance accuracy and efficiency in real time scenarios?*
RQ6: *How does the proposed $\ell_{2,1}$ norm penalty compare with the standard ℓ_2 norm?*

Datasets. We evaluate the effectiveness of `FAITH` on 11 real-world benchmark datasets: `Cora`, `Citeseer`, `Pubmed`, `A-Photo`, `Coauthor-CS`, `Physics`, `Ogbn-Arxiv`, `Squirrel`, `Chameleon`, `Credit` and `Income`. These datasets span two goals: (i) standard benchmarks for graph representation learning and GNN-to-MLP distillation, and (ii) graph datasets with sensitive attributes for group fairness evaluation. In particular, `Credit` and `Income` include demographic information, allowing evaluation of both group and individual fairness. Descriptions are in Appendix B.

Baselines and Evaluation Settings. We evaluate `FAITH` against several baselines, including GNN Teacher, **MLP**, **GLNN** (Zhang et al. 2022), **NOSMOG** (Tian et al. 2022), **KRD** (Wu et al. 2023), **VQGraph** (Yang et al. 2024), and

Dataset	Metric	MLP	SAGE	GLNN	NOSMOG	KRD	TINED	FAITH
Transductive Setting								
Cora	Acc	58.90$_{(1.43)}$	82.18$_{(0.95)}$	81.50$_{(1.21)}$	83.20$_{(0.82)}$	81.20$_{(1.07)}$	81.80$_{(1.62)}$	**83.96$_{(0.47)}$**
	IF$_{\times 10^3}$	549.90$_{(378.41)}$	194.50$_{(64.89)}$	71.13$_{(68.23)}$	104.48$_{(110.83)}$	129.30$_{(43.39)}$	171.21$_{(50.32)}$	**12.06$_{(1.30)}$**
Citeseer	Acc	60.28$_{(1.43)}$	70.86$_{(0.35)}$	71.46$_{(1.79)}$	72.70$_{(0.74)}$	72.35$_{(1.46)}$	73.03$_{(1.18)}$	**74.10$_{(0.60)}$**
	IF$_{\times 10^3}$	34.53$_{(1.38)}$	3.79$_{(2.88)}$	5.68$_{(3.85)}$	5.83$_{(2.22)}$	6.70$_{(3.90)}$	8.16$_{(6.43)}$	**1.78$_{(1.57)}$**
Pubmed	Acc	72.20$_{(0.55)}$	77.38$_{(0.81)}$	77.97$_{(3.35)}$	78.69$_{(2.90)}$	80.51$_{(0.67)}$	75.51$_{(3.22)}$	**81.53$_{(0.33)}$**
	IF$_{\times 10^3}$	2284.90$_{(988.62)}$	936.83$_{(30.35)}$	413.38$_{(42.92)}$	323.66$_{(84.07)}$	278.41$_{(27.87)}$	197.88$_{(68.86)}$	**47.96$_{(3.38)}$**
Photo	Acc	80.09$_{(2.09)}$	90.03$_{(2.05)}$	90.63$_{(0.09)}$	91.74$_{(0.25)}$	91.68$_{(2.17)}$	92.27$_{(0.41)}$	**92.48$_{(0.03)}$**
	IF$_{\times 10^5}$	263.73$_{(107.19)}$	196.08$_{(66.03)}$	131.94$_{(39.21)}$	137.49$_{(16.73)}$	126.16$_{(9.98)}$	175.74$_{(63.39)}$	**12.91$_{(1.10)}$**
CS	Acc	90.02$_{(1.05)}$	89.42$_{(0.23)}$	92.98$_{(0.21)}$	93.02$_{(0.14)}$	**93.99$_{(0.02)}$**	91.46$_{(0.42)}$	93.37$_{(0.21)}$
	IF$_{\times 10^5}$	2919.62$_{(1382.59)}$	734.13$_{(83.51)}$	1397.25$_{(889.13)}$	1176.63$_{(542.36)}$	889.42$_{(510.12)}$	984.87$_{(107.73)}$	**7.10$_{(0.24)}$**
Physics	Acc	88.23$_{(1.73)}$	91.30$_{(0.98)}$	92.20$_{(1.65)}$	93.68$_{(0.10)}$	93.58$_{(0.72)}$	92.18$_{(1.25)}$	**94.04$_{(0.13)}$**
	IF$_{\times 10^6}$	1167.06$_{(482.31)}$	43.59$_{(43.92)}$	85.39$_{(46.65)}$	21.63$_{(24.54)}$	22.39$_{(24.00)}$	25.14$_{(13.49)}$	**0.17$_{(0.02)}$**
Arxiv	Acc	54.35$_{(1.73)}$	69.74$_{(0.98)}$	69.20$_{(1.65)}$	69.45$_{(0.10)}$	69.60$_{(0.72)}$	65.65$_{(0.72)}$	**70.83$_{(0.26)}$**
	IF$_{\times 10^6}$	9508.41$_{(482.19)}$	2395.05$_{(439.23)}$	5864.06$_{(246.48)}$	4521.10$_{(245.39)}$	3129.90$_{(240.00)}$	2012.69$_{(102.61)}$	**902.11$_{(79.78)}$**
Inductive Setting								
Cora	Acc	60.76$_{(0.22)}$	**79.58$_{(0.56)}$**	72.64$_{(0.39)}$	72.92$_{(1.30)}$	73.43$_{(0.49)}$	73.05$_{(1.51)}$	73.62$_{(0.41)}$
	IF$_{\times 10^3}$	2.11$_{(0.38)}$	2.93$_{(0.01)}$	4.63$_{(0.03)}$	4.37$_{(0.40)}$	3.92$_{(0.06)}$	3.06$_{(0.61)}$	**2.03$_{(0.05)}$**
Citeseer	Acc	61.02$_{(0.69)}$	70.04$_{(0.58)}$	69.92$_{(0.83)}$	70.17$_{(1.59)}$	70.08$_{(0.92)}$	70.77$_{(1.62)}$	**71.03$_{(0.86)}$**
	IF$_{\times 10^3}$	1.69$_{(0.09)}$	0.50$_{(0.03)}$	0.78$_{(0.09)}$	0.81$_{(0.10)}$	0.76$_{(0.09)}$	0.62$_{(0.19)}$	**0.33$_{(0.02)}$**
Pubmed	Acc	72.70$_{(0.41)}$	77.68$_{(0.64)}$	78.98$_{(0.38)}$	80.03$_{(3.74)}$	80.34$_{(0.37)}$	75.28$_{(3.83)}$	**80.44$_{(0.17)}$**
	IF$_{\times 10^3}$	51.52$_{(0.76)}$	0.44$_{(0.08)}$	0.45$_{(0.06)}$	0.51$_{(0.02)}$	0.40$_{(0.04)}$	0.80$_{(0.19)}$	**0.29$_{(0.02)}$**
Photo	Acc	79.11$_{(2.40)}$	88.76$_{(1.22)}$	89.54$_{(2.20)}$	89.90$_{(1.79)}$	89.60$_{(1.83)}$	89.87$_{(1.01)}$	**90.25$_{(2.55)}$**
	IF$_{\times 10^3}$	213.73$_{(84.43)}$	138.71$_{(41.74)}$	153.03$_{(31.51)}$	90.90$_{(17.90)}$	71.71$_{(13.99)}$	120.45$_{(19.14)}$	**54.69$_{(12.28)}$**
CS	Acc	90.19$_{(1.22)}$	88.77$_{(0.84)}$	92.01$_{(0.87)}$	91.25$_{(1.76)}$	92.50$_{(1.06)}$	91.57$_{(0.60)}$	**92.74$_{(0.84)}$**
	IF$_{\times 10^3}$	3856.90$_{(1724.44)}$	118.64$_{(69.17)}$	349.70$_{(154.81)}$	743.73$_{(112.32)}$	317.74$_{(448.12)}$	629.39$_{(572.75)}$	**114.69$_{(76.95)}$**
Physics	Acc	90.12$_{(1.41)}$	91.69$_{(2.46)}$	92.29$_{(3.06)}$	93.19$_{(2.21)}$	93.74$_{(0.86)}$	93.16$_{(1.12)}$	**93.96$_{(1.20)}$**
	IF$_{\times 10^3}$	2509.75$_{(1059.59)}$	869.62$_{(359.88)}$	351.40$_{(183.41)}$	767.82$_{(344.70)}$	413.58$_{(442.96)}$	785.67$_{(341.59)}$	**99.44$_{(9.08)}$**
Arxiv	Acc	55.42$_{(0.23)}$	**71.10$_{(0.21)}$**	61.26$_{(0.13)}$	68.21$_{(0.24)}$	63.23$_{(1.05)}$	59.72$_{(0.35)}$	68.95$_{(0.81)}$
	IF$_{\times 10^3}$	310.41$_{(20.23)}$	102.09$_{(19.11)}$	129.91$_{(31.09)}$	272.65$_{(82.78)}$	240.53$_{(20.53)}$	226.43$_{(70.78)}$	**27.54$_{(7.64)}$**

Table 2: Average classification accuracy (%) and Individual Fairness (IF) with standard deviation (in subscript) across datasets in both **transductive** and **inductive** settings. The best performance is highlighted in **bold**, second-best is underlined.



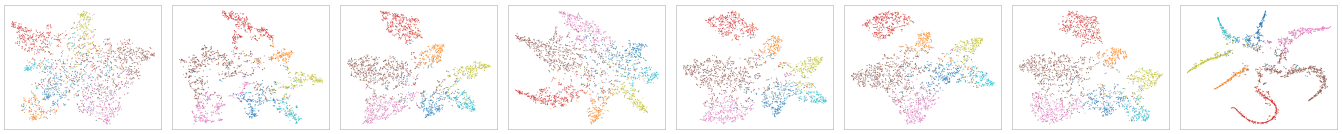
(a) Bias-Performance Coefficient (BPC)

(b) FAITH⋆ Pareto-optimal visualization

Figure 4: **(a)** Bias-performance coefficient (↑) plots for various methods on the Income and Credit datasets. Ideal methods appear near the *top-right corner*. **(b)** Pareto optimality landscape for FAITH⋆ across varying fairness regularizers γ and $\gamma^\star$ on Credit dataset. Each cell shows a parameter pair with AUC at the center and color indicating the fairness.

TINED (Zhou et al. 2025). Following (Zhang et al. 2022), we use SAGE as the default teacher, and further demonstrate the generalization of FAITH to other GNN architectures. We also compare against **FairGNN** (Dai and Wang 2021) and **NIFTY** (Agarwal, Lakkaraju, and Zitnik 2021) as fairness-aware GNN baselines. Evaluation settings, experimental details, and evaluation metrics are described in Appendix D.

Utility and Individual Fairness (RQ1). Table 2 summarizes results across both evaluation settings. FAITH consistently achieves the best trade-off between utility and fairness. In



(a) MLP (b) SAGE (c) GLNN (d) NOSMOG (e) KRD (f) TINED (g) VQGraph (h) FAITH

Figure 5: t-SNE visualization of learned representations on Cora, colors denotes different classes.

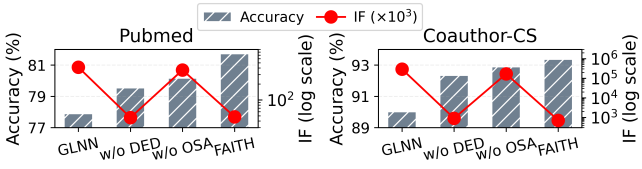


Figure 6: Ablation study of different FAITH variants.

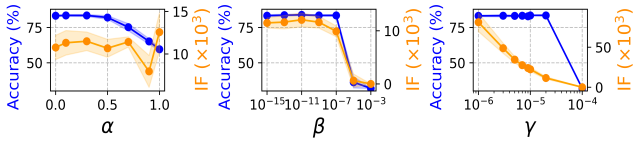


Figure 7: Hyperparameter sensitivity analysis on α, β, and γ (log scale) on the Cora dataset. IF values are in thousands.

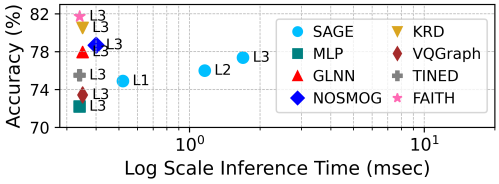


Figure 8: Accuracy vs inference time tradeoff.

the transductive setting, it delivers the highest utility on six of seven datasets and outperforms all baselines in individual fairness. In the inductive setting, FAITH delivers competitive performance compared to all distillation baselines. Although FAITH slightly underperforms KRD in utility on the Co-Author CS dataset, it achieves a significant 99.20% improvement in individual fairness. Additional discussion and results are provided in Appendix F.

Joint Individual and Group Fairness (RQ2). We jointly evaluate individual and group fairness in the transductive setting. As shown in Figure 4(a), existing distillation baselines perform poorly, exhibiting significantly lower BPC scores. FAITH achieves a substantial reduction in individual unfairness, albeit with a trade-off in group fairness. This observation aligns with prior studies, which have established an inherent conflict between individual and group fairness, where optimizing for one often leads to a deterioration in the other (Binns 2020). To address this, FAITH$\star$ introduces an additional group fairness constraint, achieving the best balance between fairness metrics while maintaining competitive utility. Notably, FAITH$\star$ outperforms both FairGNN and NIFTY, demonstrating its ability to learn fair representations even when distilled from an unfair teacher. Figure 4(b) further illustrates the Pareto frontier obtained by varying fairness regularization in FAITH$\star$, capturing the trade-offs between individual fairness, group fairness, and utility. Additional discussion is provided in Appendix F.

Different Teacher GNNs (RQ3). To show the generalizability of FAITH with different teacher networks, we report average accuracy and individual fairness on Cora, Pubmed, and A-Photo using GCN, GAT, and APPNP teachers in Appendix F. Student MLPs trained with FAITH consistently surpass other distillation methods in both utility and fairness.

Ablation Study and Hyperparameter Sensitivity (RQ4). We examine the contribution of each component in the FAITH framework using SAGE as the teacher. The DED and OSA blocks are evaluated independently. Results in Figure 6 show their positive impact under the transductive setting. The effect of the group aware fairness term in FAITH$\star$ is shown in Figure 4. We also analyze hyperparameter sensitivity in Figure 7. These results show that FAITH achieves stable performance across a range of hyperparameters.

Accuracy vs. Inference Time (RQ5). Figure 8 shows the tradeoff between inference time and accuracy on the Pubmed dataset. A 3 layer SAGE teacher is used to distill 3 layer MLPs with different methods. Baselines with 1, 2, and 3 layer SAGE models are included for comparison. Labels L1, L2 and L3 indicate the number of layers. FAITH gives the best accuracy among distilled models while keeping inference time close to shallow MLPs.

ℓ_2 vs. $\ell_{2,1}$ Penalty (RQ6). Refer to Apendix F for details. Results show that FAITH with the $\ell_{2,1}$ norm achieves better sparsity and a stronger tradeoff between utility and fairness.

Visualization. Figure 5 shows t-SNE (Maaten and Hinton 2008) visualizations of all representations. FAITH yields more compact representations. Visualization results on additional datasets are provided in Appendix F.

Conclusion

In this paper, we introduced FAITH, a comprehensive framework for efficient and fairness-aware inference on graph datasets via GNN-to-MLP distillation. To align MLPs with the graph signal denoising principles of the teacher network, we proposed node-level Dirichlet energy distillation. To enforce individual fairness, we introduced an $\ell_{2,1}$-norm-based oracle similarity preserving objective, regulating the Lipschitz behavior of the network. The framework was further extended to incorporate group-level fairness. Benchmark results show that FAITH balances utility and fairness effectively.

Acknowledgments

Vipul Kumar Singh is supported by the Prime Minister's Research Fellowship (Grant No. 1402107). Jyotismita Barman is supported by the Tata Consultancy Services Research Scholar Program. Sandeep Kumar is supported by the DST INSPIRE Faculty Grant (Grant No. MI02322G) and Biofin Capital. This work is also supported by the Science and Engineering Research Board, India (Grant No. RP04820G). The authors thank the anonymous reviewers for their thoughtful and constructive comments.

References

Agarwal, C.; Lakkaraju, H.; and Zitnik, M. 2021. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, 2114–2124. PMLR.

Binns, R. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.

Dai, E.; and Wang, S. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM international conference on web search and data mining*, 680–688.

Dong, Y.; Ma, J.; Wang, S.; Chen, C.; and Li, J. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10): 10583–10602.

Ghosh, B.; Basu, D.; and Meel, K. S. 2022. Algorithmic fairness verification with graphical models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9539–9548.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Kang, J.; He, J.; Maciejewski, R.; and Tong, H. 2020. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 379–389.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liang, C.; Zuo, S.; Zhang, Q.; He, P.; Chen, W.; and Zhao, T. 2023. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, 20852–20867. PMLR.

Liu, X.; Yan, M.; Deng, L.; Li, G.; Ye, X.; Fan, D.; Pan, S.; and Xie, Y. 2022. Survey on Graph Neural Network Acceleration: An Algorithmic Perspective. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5521–5529. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Liu, Z.; Wan, G.; Prakash, B. A.; Lau, M. S.; and Jin, W. 2024. A review of graph neural networks in epidemic modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6577–6587.

Ma, Y.; Liu, X.; Zhao, T.; Liu, Y.; Tang, J.; and Shah, N. 2021. A unified view on graph neural networks as graph signal denoising. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1202–1211.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; et al. 2022. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1): 93.

Rosenblatt, L.; and Witter, R. T. 2023. Counterfactual fairness is basically demographic parity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14461–14469.

Rue, H.; and Held, L. 2005. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Singh, V. K.; Kumar, S.; Prasad, A.; et al. 2024. A Unified Optimization-Based Framework for Certifiably Robust and Fair Graph Neural Networks. *IEEE Transactions on Signal Processing*.

Song, W.; Dong, Y.; Liu, N.; and Li, J. 2022. Guide: Group equality informed individual fairness in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1625–1634.

Tian, Y.; Zhang, C.; Guo, Z.; Zhang, X.; and Chawla, N. 2022. Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency. In *The Eleventh International Conference on Learning Representations*.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. Ieee.

Wang, Y.-X.; Sharpnack, J.; Smola, A. J.; and Tibshirani, R. J. 2016. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105): 1–41.

Wu, L.; Lin, H.; Huang, Y.; and Li, S. Z. 2023. Quantifying the knowledge in gnns for reliable distillation into mlps. In *International Conference on Machine Learning*, 37571–37581. PMLR.

Yang, L.; Tian, Y.; Xu, M.; Liu, Z.; Hong, S.; Qu, W.; Zhang, W.; CUI, B.; Zhang, M.; and Leskovec, J. 2024. VQGraph: Rethinking Graph Representation Space for Bridging GNNs and MLPs. In *The Twelfth International Conference on Learning Representations*.

Zhang, S.; Liu, Y.; Sun, Y.; and Shah, N. 2022. Graph-less Neural Networks: Teaching Old MLPs New Tricks via Distillation. In *International Conference on Learning Representations*.

Zhang, X.-M.; Liang, L.; Liu, L.; and Tang, M.-J. 2021. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12: 690049.

Zhou, Y.; King, T.; Zhao, H.; Huang, Y.; Riedel, T.; and Beigl, M. 2024. MLP-HAR: Boosting Performance and Efficiency of HAR Models on Edge Devices with Purely Fully Connected Layers. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, 133–139.

Zhou, Z.; Ding, Z.; Shi, J.; Qing, L.; and Shen, S. 2025. TINED: GNNs-to-MLPs by Teacher Injection and Dirichlet Energy Distillation. In *Forty-second International Conference on Machine Learning*.