Beyond Sparse Benchmarks: Evaluating GNNs with Realistic Missing Features

Francesco Ferrini
University of Trento
Trento, Italy
francesco.ferrini@unitn.it

Veronica Lachi
Fondazione Bruno Kessler
Trento, Italy
vlachi@fbk.eu

Antonio Longa
University of Trento
Trento, Italy
antonio.longa@unitn.it

Bruno Lepri
Fondazione Bruno Kessler
Trento, Italy
lepri@fbk.eu

Andrea Passerini
University of Trento
Trento, Italy
andrea.passerini@unitn.it

Xin Liu
National Institute of Advanced Industrial Science and Technology
Tokyo, Japan
xin.liu@aist.go.jp

Manfred Jaeger
Aalborg University
Aalborg, Denmark
jaeger@cs.aau.dk

Abstract

Handling missing node features is a critical challenge for deploying Graph Neural Networks (GNNs) in real-world applications such as healthcare and sensor networks. This has led to a number of recent works exploring techniques for learning GNNs from incomplete data. However, existing evaluations are often based on benchmark datasets with high-dimensional but very sparse node features, where predictive performance degrades only slowly as the proportion of missing values increases. In this paper we move towards more challenging and realistic scenarios by considering datasets in which the predictive signal is more sensitive to feature incompleteness. We provide a theoretical background for clearly identifying relevant assumptions on the missingness mechanism, and for analyzing their implications for different solution approaches. Based on this analysis, we introduce the `GNNmim` approach for node classification in graphs with incomplete feature data. Experiments show that `GNNmim` consistently outperforms more complex models across a range of datasets and levels of missingness.

1 Introduction

Missing features are a pervasive challenge in many real-world machine learning applications, such as healthcare [4, 11], IoT sensor networks [8, 12, 1], and database migration [2, 13]. This issue naturally extends to Graph Neural Networks (GNNs), which are increasingly applied in domains where incomplete data is common. Here, we focus specifically on the problem of *missing node features*, a setting that has received growing attention in the GNN literature.

Several methods have been proposed to address this challenge, ranging from simple mean imputation to sophisticated architectures that perform joint imputation and prediction during training. These are typically evaluated by synthetically removing features from standard benchmarks such as CORA, CITESEER, and PUBMED. Yet these datasets rely on extremely sparse bag-of-words features, raising

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

a crucial question: *Can we meaningfully assess robustness to missing features when most entries are already zero?* Our theoretical analysis shows that in highly sparse settings, the information loss from additional missingness is negligible until corruption becomes extreme. Empirically, existing methods maintain high accuracy until more than 90% of entries are missing, limiting the ability of such benchmarks to differentiate between approaches.

To move beyond this limitation, we identify and propose a set of datasets, one synthetic and three real-world, characterized by dense, raw features. These datasets provide a more realistic testbed, with low sparsity, high informativeness, and both *feature-structure complementarity* and *separability* [5].We also revisit the design of the missingness mechanism. In addition to the widely adopted *uniform randomly missingness* and the *structurally missingness* settings from prior work, we introduce a novel *correlation-driven MCAR* mechanism, where missingness probabilities are correlated with node labels, offering a more challenging evaluation.

Finally, we introduce `GNNmim`, a simple method that augments the node feature matrix with a binary mask indicating missing values, processed by a standard GNN without learned imputation. Despite its simplicity, `GNNmim` consistently outperforms more complex models across datasets and missingness settings.

Contributions. Our main contributions are:

1. A theoretical analysis showing that the effect of missingness strongly depends on feature sparsity, with an information-theoretic bound on the incurred loss.
2. A set of dense, informative datasets (one synthetic, three real-world) offering a more suitable testbed for GNNs under missingness.
3. An evaluation covering three MCAR mechanisms: uniform randomly missingness, structurally missingness, and our proposed correlation-driven MCAR.
4. The `GNNmim` method, which outperforms more complex models while highlighting limitations of current evaluation practices.

2 Learning from Incomplete Graph Data

We consider an attributed graph $G = (V, E, \mathbf{X}, \mathbf{Y})$, where $V = \{1, \ldots, n\}$ is the set of nodes, $E \subseteq V \times V$ is the set of edges represented by the adjacency matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the node feature matrix with entry X_{ij} denoting feature j of node i, and $\mathbf{Y} \in \mathcal{Y}^n$ is the vector of node labels.

When data is incomplete, then some entries of $\mathbf{X}$ are unobserved. Let $\mathbf{M} \in \{0,1\}^{n \times d}$ be the missingness indicator matrix that has $M_{ij} = 1$ if x_{ij} is missing and 0 otherwise. Let $\mathbf{X}^{obs}$ be the elements of $\mathbf{X}$ for which $M_{ij} = 1$, and $\mathbf{X}^{miss}$ the elements for which $M_{ij} = 0$. The observed data from which we learn then can be written as $\mathbf{X}^{obs}, \mathbf{Y}, \mathbf{M}$. We note that we here make the assumption that $\mathbf{Y}$ is fully observed in the (training) data, and that there is no uncertainty about the graph structure E. The distribution of the data then can be parameterized as

$$P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{X}^{obs}, \mathbf{Y}, \mathbf{M}) = \int_{\mathbf{X}^{miss}} P_{\boldsymbol{\theta}}(\mathbf{X}) P_{\boldsymbol{\gamma}}(\mathbf{Y}|\mathbf{X}) P_{\boldsymbol{\lambda}}(\mathbf{M}|\mathbf{X}, \mathbf{Y}), \tag{1}$$

where $\mathbf{X} = \mathbf{X}^{obs} \cup \mathbf{X}^{miss}$, $P_{\boldsymbol{\theta}}$ is the node feature distribution, $P_{\boldsymbol{\gamma}}$ is the conditional label distribution, and $P_{\boldsymbol{\lambda}}$ represents the *missingness mechanism*. Though not explicitly reflected in the notation, all these distributions will usually depend on the underlying graph structure, which will typically induce dependencies among the rows of $\mathbf{X}$, and among the elements of $\mathbf{Y}$.

A GNN for node classification with complete feature data is a model $P_{\boldsymbol{\gamma}}(\mathbf{Y}|\mathbf{X})$ with $\boldsymbol{\gamma}$ the weights of the GNN. For classification, we need to learn the conditional model

$$P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{X}^{obs}, \mathbf{M}) = \int_{\mathbf{X}^{miss}} P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{M}) P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{X}^{miss}|\mathbf{X}^{obs}, \boldsymbol{M}). \tag{2}$$

The classical *missing (completely) at random (M(C)AR)* assumptions [15] simplify this problem. The original M(C)AR assumptions have been formulated in the context of estimating the parameter of a generative distribution. It has been observed that more specialized variations of the original

definitions can be more pertinent in the context of classification [6, 9]. In the following we give formulations of M(C)AR for classification that provide the foundations for our theoretical analysis.

Definition 1. The joint distribution $P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}$ is *feature*-MAR, if

$$P_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{M}|\mathbf{X}^{miss}, \mathbf{X}^{obs}) = P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{M}|\mathbf{X}^{obs}). \tag{3}$$

It is *label*-MAR if

$$P_{\boldsymbol{\lambda}}(\mathbf{M}|\boldsymbol{X}, \boldsymbol{Y}) = P_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{M}|\boldsymbol{X}). \tag{4}$$

The distribution is MCAR, if

$$P_{\boldsymbol{\lambda}}(\mathbf{M}|\boldsymbol{X}, \boldsymbol{Y}) = P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{M}). \tag{5}$$

In (3)-(5) all probability functions are indexed with the parameters they actually depend on. Note, for example, that the conditional of $\mathbf{M}$ given $\mathbf{X}$ requires marginalization over $\mathbf{Y}$, and thereby also depends on the parameter $\boldsymbol{\gamma}$. MCAR implies both feature- and label-MAR.

The simplest realization of an MCAR mechanism is *uniform independent missingness* in which entries of $\mathbf{X}$ are missing with a fixed missingness probability μ. This can be generalized by defining a missingness probability matrix $\boldsymbol{\mu} \in [0, 1]^{n\times d}$ specifying potentially different missingness probabilities for different entries of $\mathbf{X}$. Another missingness mechanism often considered in the graph learning literature is *structural missingness* where randomly selected rows of $\mathbf{X}$ are set to missing. This, too, is still an MCAR mechanism, but now with internal dependencies among the components of $\mathbf{M}$.

MAR assumptions allow us to eliminate the coarsening model $P_{\boldsymbol{\lambda}}$ from (2). The following proposition states this classical *ignorability* result in a version most suitable in our context.

Theorem 1. If $P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}$ is feature-MAR and label-MAR, then (2) simplifies to

$$\int_{\mathbf{X}^{miss}} P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X}) P_{\boldsymbol{\theta}}(\mathbf{X}^{miss}|\mathbf{X}^{obs}). \tag{6}$$

The proof is straightforward by rewriting the two factors on the right of (2) using Bayes's rule, and plugging in (3) and (4). Apart from eliminating the missingness mechanism, (6) also simplifies (2) by separating the marginal feature model $P_{\boldsymbol{\theta}}$, and the conditional label distribution $P_{\boldsymbol{\gamma}}$. Formulation (6) still poses two major challenges: it requires a feature distribution model $P_{\boldsymbol{\theta}}$ when in reality we only are interested in the conditional model $P_{\boldsymbol{\gamma}}$, and the integration over $\mathbf{X}^{miss}$ is usually intractable.

The simplest approach to address these problems is to approximate the integral (6) by evaluating $P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X})$ at a single imputed value $\mathbf{X} = \textit{impute}(\mathbf{X}^{miss})$. This does not require an explicit model for $P_{\boldsymbol{\theta}}$, but relies on the implicit assumption that the imputed value $\textit{impute}(\mathbf{X}^{miss})$ has high probability under $P_{\boldsymbol{\theta}}$. A simple example is *mean-imputation*, in which missing values of a given feature are filled with the mean of that feature. Similarly, `PCFI` [17] does not require an explicit model for $P_{\boldsymbol{\theta}}$; it introduces a confidence-guided imputation scheme where pseudo-confidence is derived from the shortest-path distance to observed features, and combines channel-wise diffusion with inter-channel propagation to recover a single estimate of $\mathbf{X}$. `GOODIE` [19] approximates the integral in (6) using a combination of label propagation and `FP` [14], which propagates features by minimizing a Dirichlet energy function, whereas `FairAC` [10] does so by aggregating, via an attention mechanism, the representations from neighbors of nodes with missing features.

Other methods explicitly model $P_{\boldsymbol{\theta}}$. The `GCNmf` approach of Taguchi et al. [16] introduces a model of $P_{\boldsymbol{\theta}}$ in the form of a mixture of Gaussians, and approximates (6) by $P_{\boldsymbol{\gamma}}(\boldsymbol{Y}, |, \mathbb{E}_{\theta}[\mathbf{L}1 \mid \mathbf{X}^{obs}])$, where $\mathbb{E}_{\theta}[\mathbf{L}_1 \mid \mathbf{X}^{obs}]$ is the expected activation at the first layer of the GNN defining $P_{\boldsymbol{\gamma}}$. Finally, `GSPN`[7] explicitly models $P_{\boldsymbol{\theta}}$ with graph-induced sum–product networks, so missing features are handled by exact marginalization.

An alternative to all these approaches that work entirely with models $P_{\boldsymbol{\theta}}, P_{\boldsymbol{\gamma}}$ for the (complete) data distribution is to include the missingness mechanism explicitly in a model $P_{\boldsymbol{\gamma}^+}(\boldsymbol{Y}|\mathbf{X}^{obs}, \boldsymbol{M})$, that directly captures the left side of (2). We here write $\boldsymbol{\gamma}^+$ for the parameters of the model to emphasize that it can be structurally similar to a model $P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X})$, but different in that it has the missingness matrix $\boldsymbol{M}$ as an explicit extra input.

This modeling strategy, often referred to as the Missing Indicator Method (MIM), has been studied in the context of supervised learning with missing features [18], but, to the best of our knowledge, it has not been explored in the context of graph machine learning. In this work, we propose a GNN-based

instantiation of the MIM framework, which we call `GNNmim`. in `GNNmim`, we implement P_{γ^+} as a GNN, we construct the matrix *zero-pad*($\mathbf{X}^{obs}$) in which missing values are filled in by zeros, and use the concatenation *zero-pad*($\mathbf{X}^{obs}$)$[i,:]||\boldsymbol{M}[i,:]$ as the feature vector for node i in an otherwise standard GNN architecture[1]. `GNNmim` does not rely on any MAR assumptions, and thereby can be expected to perform more robustly than other approaches under different missingness mechanisms. As our experiments in Section 4 show, this simple yet principled strategy yields robust performance across a wide variety of missingness scenarios.

3 Are we evaluating GNNs for missing data on the correct dataset?

A rigorous evaluation of GNNs under feature missingness requires not only suitable models but also appropriate datasets. Recent work has stressed the importance of dataset choice in benchmarking. The position paper Bechler-Speicher et al. [3] argues that current practices in benchmark selection need substantial revision, while Coupette et al. [5] propose two diagnostics, *performance separability* and *mode complementarity*, to assess dataset informativeness. In the context of missing node features, dataset suitability is even more critical: models should be tested where missingness meaningfully affects performance and reasoning under missingness is non-trivial. Nevertheless, current practice still relies on benchmarks such as CORA, CITESEER, PUBMED, AMAZONCOMPUTERS, and AMAZONPHOTO.

In these datasets, node features are constructed as follows: CORA, CITESEER and PUBMED use binary bag-of-words features, while AMAZONCOMPUTERS and AMAZONPHOTO use TF-IDF vectors. These feature matrices are typically very sparse, which we quantify using the notion of *feature sparsity*, formally defined as:

Definition 2 (Feature Sparsity)**.** Given a node feature matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the *feature sparsity* is defined as the proportion of zero entries: $s(\mathbf{X}) = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbf{1}[X_{ij}=0]$, where $\mathbf{1}[\cdot]$ denotes the indicator function.

Table 1: Feature sparsity across benchmarks and custom datasets.

Dataset	Sparsity ↓	Features
CORA	0.9873	BoW
PUBMED	0.8998	BoW
CITESEER	0.9915	BoW
AMAZONPHOTO	0.6526	TF-IDF
AMAZONCOMPUTERS	0.6516	TF-IDF
SYNTHETIC	0.0000	Gaussian
AIR	0.1615	Raw
ELECTRIC	0.2000	Raw
TADPOLE	0.0000	Raw

The sparsity values of the benchmark datasets are reported in Table 1 (first five rows). All datasets exhibit substantial sparsity, with more than 50% of features being zero across all the datasets, with Citeseer reaching an extreme sparsity level of approximately 99%. This raises a crucial question: does it make sense to evaluate models designed to handle missing features on datasets where the feature representations are already extremely sparse? In such sparse settings, a high probability of missingness is needed to induce a meaningful information loss. Otherwise, the observed model performance under missingness may reflect artifacts of the dataset rather than the robustness of the method. We formalize this observation in the following theorem.

Theorem 2. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} \in \mathcal{Y}^n$ be random variables, $\mathbf{M} \in \{0,1\}^{n\times d}$ be a missingness mask and $\mathbf{X}^{obs}$ denote the observed (incomplete) data. We encode the pair $(\mathbf{X}^{obs}, \mathbf{M})$ with the random variable $\tilde{\mathbf{X}}$ with

$$\tilde{X}_{ij} \;=\; \begin{cases} X_{ij}, & M_{ij}=1, \\ ?, & M_{ij}=0. \end{cases}$$

Let the change in the information be defines as $\Delta \;:=\; I(\mathbf{Y};\tilde{\mathbf{X}}) \;-\; I(\mathbf{Y};\mathbf{X})$. Then,

1. If the missingness is label-MAR, then $\Delta \;\leq\; 0$.
2. If $\mathbf{X} \in \{0,1\}^{n\times d}$ and the missingness is uniform MCAR, independently of $(\mathbf{X},\mathbf{Y})$ and identically over (i,j), being the (random) sample sparsity $s(\mathbf{X})$ be defined as in Definition

[1]We deliberately here say "zero-padding" rather than "zero-imputation". The latter would imply that we view the zeros as somehow reasonable stand-ins for the true unobserved values. We view the zeros as arbitrary placeholders. Ideally, the trained model will learn to ignore these values when the corresponding missingness indicator is 1.

2, then

$$-\, nd\,\mu\, h_2\big(\mathbb{E}[s(\mathbf{X})]\big) \;\leq\; \Delta \;\leq\; 0,$$

where $h_2(u) = -u \log u - (1-u)\log(1-u)$.

The proof can be found in Appendix A. Theorem 2 demonstrates that when feature sparsity is high, a very large amount of missingness is required to produce a meaningful loss of information. This theoretical insight is also reflected empirically. As shown in Figure 1, which reports the case of Cora, Citeseer and PubMed under *uniform random* (UR) missingness, GNN-based models maintain consistently high F1 scores across almost all levels of missingness, with a noticeable drop only beyond 90%.

These theoretical and empirical results, confirms that such benchmarks do not meaningfully differentiate between approaches, casting doubt on their suitability for evaluating GNNs under feature missingness. As a consequence, we argue for the use of datasets where missingness poses a real and measurable challenge. To support this perspective, we introduce a set of alternative datasets, both synthetic and real-world, that are better aligned with the characteristics required to meaningfully evaluate GNNs under feature missingness. These datasets aim to open a new direction for research by highlighting the limitations of existing benchmarks and enabling more principled empirical analysis.

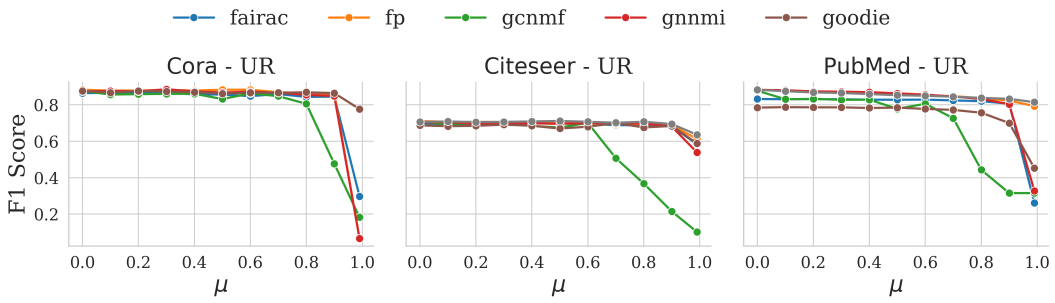

Figure 1: F1 scores on Cora, Citeseer and PubMed under uniform MCAR.

(1) Synthetic dataset with controlled missingness. We design a Barabási–Albert graph with Gaussian node features and labels generated by a fixed two-layer GCN trained on complete features. This ensures that, without missingness, high classification accuracy is attainable. The dataset thus serves as a clean testbed to isolate the effect of feature sparsity under a well-defined ground truth.

(2) Real-world datasets with meaningful features. We further consider real datasets where node features correspond to observable properties: 1) **AIR**, an IoT sensor network with environmental measurements and sensor status labels; 2) **ELECTRIC**, an electrical sensor graph with real-valued features and operational condition labels; 3) **TADPOLE**, a medical dataset where nodes are patients, features include clinical and imaging biomarkers, and labels correspond to diagnoses.

Table 2: Evaluation of P1 (*performance separability*) and P2 (*mode complementarity*) on our custom datasets. Each cell reports the KS statistic and associated p-value for separability under six perturbation settings. $\gamma_{1,1}$ indicates the feature-structure complementarity. Datasets satisfying each property (as per Coupette et al. [5]) are marked with ✓.

Dataset	Empty Feat.	Random Feat.	Complete Feat.	Empty Graph	Random Graph	Complete Graph	$\gamma_{1,1}$	P1	P2
SYNTHETIC	1.00 (8.80e-62)	1.00 (8.80e-62)	1.00 (1.93e-14)	1.00 (1.03e-17)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.62	✓	✓
AIR	1.00 (8.80e-62)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.67 (1.53e-30)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.68	✓	✓
ELECTRIC	1.00 (8.80e-62)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.98 (1.90e-57)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.69	✓	✓
TADPOLE	1.00 (8.80e-62)	0.90 (5.31e-44)	0.61 (4.22e-18)	0.77 (1.53e-30)	1.00 (8.80e-62)	1.00 (8.80e-62)	0.64	✓	✓

Both the synthetic and real-world datasets show low feature sparsity (Table 1), a necessary but not sufficient condition for benchmarking under missingness. Following the RINGS framework [5], we further evaluate two criteria: *performance separability*, measuring whether features and structure individually carry task-relevant information, and *mode complementarity*, quantifying their alignment in a task-agnostic space. Values of $\gamma_{1,1}$ above 0.5 indicate satisfactory quality, which holds for all our datasets. Those values are shown in Table 2.

4 Experiments

We conduct experiments on node classification tasks using the datasets introduced in Section 3 and the MCAR-based missingness mechanisms defined in the next paragraph. The goal is to evaluate how well existing methods and our proposed approach perform under different conditions of missing node features. We compare a range of GNN-based models specifically designed to handle missing features, including `GNNmi`, `GCNmf`, `GOODIE`, `GSPN`, `PCFI`, `FP`, and `FairAC`, introduces in Section 2, alongside our proposed method, `GNNmim` (Details in Appendix C). Code available at[2].

Missingness mechanisms Most prior works adopt a masking scheme under the MCAR assumption (Definition 1). The common variant is ***Uniform Randomly MCAR (UR)***, where each feature entry is masked independently with probability $\mu \in [0, 1]$. Another variant, used for example by Taguchi et al. [16], Um et al. [17], is ***Structural MCAR (S)***, where entire feature vectors of randomly selected nodes are masked. We additionally introduce a more challenging mechanism, ***Correlation–Dependent MCAR (CD)***, which masks features at the column level with probability proportional to their mutual information with the label. Each entry X_{ij} is masked independently with $P(M_{ij} = 1) = \rho \cdot w_j$, where $w_j \propto I(X_j; Y)$ and $\rho \in [0, 1]$ is scaled to match the target missingness rate.

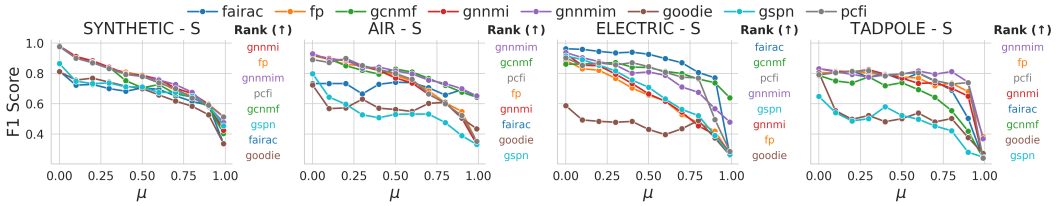

Figure 2: F1 score vs. feature missingness (μ) on our datasets under *Structural MCAR* (S). To the right, models are ranked by mean F1 across all μ values (best on top). Complete plots in App. B.

Results We evaluate performance across missingness rates in {0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 99%}. Figure 2 shows results for all datasets and the *Structural MCAR* mechanism (complete plots for all missingness mechanisms are reported in B.). Compared to results on standard benchmarks (Figure 1), models exhibit a much stronger drop in performance, even at moderate levels of missingness. On classical datasets such as Cora or Citeseer, models often appear robust until very high rates (80–90%), but as argued in Section 3, this is largely due to feature sparsity. In contrast, the proposed datasets expose more realistic and challenging behavior: performance degrades earlier and more substantially. To assess robustness, for each plot in Figure 2 we report the average model rank across missingness levels. Table 3 summarizes the overall ranks across datasets and mechanisms, with `GNNmim` emerging as the most robust method, confirming the value of explicitly leveraging missingness through a simple, assumption-free mechanism.

Table 3: Overall average rank (lower is better), obtained by aggregating the per-panel ranks in Fig. 2 across all datasets and missingness mechanisms (see Appendix B).

Model	Avg. Rank
`GOODIE`	5.50
`GSPN`	7.25
`FairAC`	4.50
`GCNmf`	4.00
`FP`	4.25
`PCFI`	3.75
`GNNmi`	3.70
`GNNmim`	**1.75**

Overall, these experiments highlight that our proposed datasets, combined with our MCAR mechanisms, expose meaningful robustness differences between methods. In particular, `GNNmim`'s consistent performance suggests that explicitly modeling missingness through simple indicators can be more effective than complex imputation strategies or probabilistic models.

5 Conclusion and Future Work

In this work, we revisited the evaluation of Graph Neural Networks under feature missingness. We showed that widely used benchmarks, dominated by highly sparse features, fail to expose the real

[2] `https://anonymous.4open.science/r/gnnmim-257C/`

challenges of incomplete data. To address this, we introduced denser and more informative datasets, and proposed `GNNmim`, a simple method that augments node features with explicit missingness indicators. Across datasets and masking mechanisms, `GNNmim` consistently matches or outperforms more complex methods. This indicates that current approaches are not necessarily superior to simple strategies when tested in realistic settings, opening the way for more principled solutions.

Future Work. Our study focused on MCAR mechanisms, but extending the evaluation to **Missing Not At Random (MNAR)** scenarios, where missingness depends on feature values, would be highly relevant in domains such as healthcare. Another important direction is to study **distribution shifts between training and test missingness**, where type or intensity differ across phases, a setting closer to real-world deployments.

Acknowledgments

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO. We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU. This work was also supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007)

References

[1] Benjamin Agbo, Hussain Al-Aqrabi, Richard Hill, and Tariq Alsboui. Missing data imputation in the internet of things sensor networks. *Future Internet*, 14(5):143, 2022.

[2] Otmane Azeroual and Meena Jha. Without data quality, there is no data migration. *Big Data and Cognitive Computing*, 5(2):24, 2021.

[3] Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M Bronstein, Mathias Niepert, Bryan Perozzi, et al. Position: Graph learning will lose relevance due to poor benchmarks. *arXiv preprint arXiv:2502.14546*, 2025.

[4] Carlijn IR Braem, Utku S Yavuz, Hermie J Hermens, and Peter H Veltink. Missing data statistics provide causal insights into data loss in diabetes health monitoring by wearable sensors. *Sensors*, 24(5):1526, 2024.

[5] Corinna Coupette, Jeremy Wayland, Emily Simons, and Bastian Rieck. No metric to rule them all: Toward principled evaluations of graph-learning datasets. *arXiv preprint arXiv:2502.02379*, 2025.

[6] Yufeng Ding and Jeffrey S Simonoff. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11(1), 2010.

[7] Federico Errica and Mathias Niepert. Tractable probabilistic graph representation learning with graph-induced sum-product networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=h7nOCxFsPg`.

[8] Rahmat Nur Faizin, Mardhani Riasetiawan, and Ahmad Ashari. A review of missing sensor data imputation methods. In *2019 5th International Conference on Science and Technology (ICST)*, volume 1, pages 1–6. IEEE, 2019.

[9] Amirata Ghorbani and James Y Zou. Embedding for informative missingness: Deep learning with incomplete data. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 437–445. IEEE, 2018.

[10] Dongliang Guo, Zhixuan Chu, and Sheng Li. Fair attribute completion on graph with missing attributes. *arXiv preprint arXiv:2302.12977*, 2023.

[11] Eugenij Moiseevich Mirkes, Timothy J Coats, Jeremy Levesley, and Alexander N Gorban. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in biology and medicine*, 75:203–216, 2016.

[12] Nwamaka U Okafor and Declan T Delaney. Missing data imputation on iot sensor networks: Implications for on-site sensor calibration. *IEEE Sensors journal*, 21(20):22833–22845, 2021.

[13] Andrei Rogers, Jani Little, and James Raymer. *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*, volume 26. Springer Science & Business Media, 2010.

[14] Emanuele Rossi, Henry Kenlay, Maria I Gorinova, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Learning on graphs conference*, pages 11–1. PMLR, 2022.

[15] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[16] Hibiki Taguchi, Xin Liu, and Tsuyoshi Murata. Graph convolutional networks for graphs containing missing features. *Future Generation Computer Systems*, 117:155–168, 2021.

[17] Daeho Um, Jiwoong Park, Seulki Park, and Jin young Choi. Confidence-based feature imputation for graphs with partially known features. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=YPKBIILy-Kt`.

[18] Mike Van Ness, Tomas M Bosschieter, Roberto Halpin-Gregorio, and Madeleine Udell. The missing indicator method: From low to high dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5004–5015, 2023.

[19] Sukwon Yun, Xin Liu, Yunhak Oh, Junseok Lee, Tianlong Chen, Tsuyoshi Murata, and Chanyoung Park. Oldie but goodie: Re-illuminating label propagation on graphs with partially observed features, 2024. URL `https://openreview.net/forum?id=TlFDFKyEIQ`.

A Proofs

Theorem 1. If $P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}$ is feature-MAR and label-MAR, then (2) simplifies to

$$\int_{\mathbf{X}^{miss}} P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X}) P_{\boldsymbol{\theta}}(\mathbf{X}^{miss}|\mathbf{X}^{obs}). \tag{6}$$

Proof.

$$P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{M}) = P_{\boldsymbol{\lambda}}(\boldsymbol{M}|\boldsymbol{X},\boldsymbol{Y}) \frac{P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X})}{P_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{M}|\boldsymbol{X})} \overset{(4)}{=} P_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\boldsymbol{X})$$

$$P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\mathbf{X}^{miss}|\mathbf{X}^{obs},\boldsymbol{M}) = P_{\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{M}|\mathbf{X}^{obs},\mathbf{X}^{miss}) \frac{P_{\boldsymbol{\theta}}(\mathbf{X}^{miss}|\mathbf{X}^{obs})}{P_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\lambda}}(\boldsymbol{M}|\mathbf{X}^{obs})} \overset{(3)}{=} P_{\boldsymbol{\theta}}(\mathbf{X}^{miss}|\mathbf{X}^{obs})$$

□

Theorem 2. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} \in \mathcal{Y}^n$ be random variables, $\mathbf{M} \in \{0,1\}^{n\times d}$ be a missingness mask and $\mathbf{X}^{obs}$ denote the observed (incomplete) data. We encode the pair $(\mathbf{X}^{obs}, \mathbf{M})$ with the random variable $\tilde{\mathbf{X}}$ with

$$\tilde{X}_{ij} \;=\; \begin{cases} X_{ij}, & M_{ij} = 1, \\ ?, & M_{ij} = 0. \end{cases}$$

Let the change in the information be defines as $\Delta \;:=\; I(\mathbf{Y};\tilde{\mathbf{X}}) \;-\; I(\mathbf{Y};\mathbf{X})$. Then,

1. If the missingness is label-MAR, then $\Delta \;\leq\; 0$.
2. If $\mathbf{X} \in \{0,1\}^{n\times d}$ and the missingness is uniform MCAR, independently of $(\mathbf{X},\mathbf{Y})$ and identically over (i,j), being the (random) sample sparsity $s(\mathbf{X})$ be defined as in Definition 2, then
$$-\; nd\,\mu\, h_2\big(\mathbb{E}[s(\mathbf{X})]\big) \;\leq\; \Delta \;\leq\; 0,$$
where $h_2(u) = -u\log u - (1-u)\log(1-u)$.

Proof. By construction $\tilde{\mathbf{X}} = g(\mathbf{X},\mathbf{M})$ for some measurable g. Thus $(\mathbf{Y}) \to (\mathbf{X},\mathbf{M}) \to \tilde{\mathbf{X}}$ is a Markov chain, and the data–processing inequality implies

$$I(\mathbf{Y};\tilde{\mathbf{X}}) \;\leq\; I(\mathbf{Y};\mathbf{X},\mathbf{M}). \tag{7}$$

Moreover, for any three random elements (A,B,C) we have the chain–rule identities

$$I(A;B,C) = I(A;C) + I(A;B \mid C). \tag{8}$$

(1) Label-MAR $\Delta \leq 0$. Assume label-MAR: $\mathbb{P}(\mathbf{M} \mid \mathbf{X},\mathbf{Y}) = \mathbb{P}(\mathbf{M} \mid \mathbf{X})$, which is equivalent to $\mathbf{Y} \perp \mathbf{M} \mid \mathbf{X}$. Applying (8) with $(A,B,C) = (\mathbf{Y},\mathbf{X},\mathbf{M})$,

$$I(\mathbf{Y};\mathbf{X},\mathbf{M}) \;=\; I(\mathbf{Y};\mathbf{X}) + I(\mathbf{Y};\mathbf{M} \mid \mathbf{X}).$$

Under label-MAR, $I(\mathbf{Y}; \mathbf{M} \mid \mathbf{X}) = 0$, hence

$$I(\mathbf{Y}; \mathbf{X}, \mathbf{M}) \;=\; I(\mathbf{Y}; \mathbf{X}). \tag{9}$$

Combining (7) and (9) yields

$$I(\mathbf{Y}; \tilde{\mathbf{X}}) \;\le\; I(\mathbf{Y}; \mathbf{X}) \quad\Longleftrightarrow\quad \Delta \;=\; I(\mathbf{Y}; \tilde{\mathbf{X}}) - I(\mathbf{Y}; \mathbf{X}) \;\le\; 0.$$

(2) Two-sided bound under uniform MCAR and α-β sparsity. Assume uniform MCAR: $M_{ij} \sim \mathrm{Bernoulli}(1-\mu)$ independently of $(\mathbf{X}, \mathbf{Y})$ and i.i.d. across (i, j), and that $\mathbb{P}\big(s(\mathbf{X}) \ge \alpha\big) \ge \beta$, where $s(\mathbf{X}) = \frac{1}{nd} \sum_{i,j} \mathbb{I}\{X_{ij} = 0\}$.

Upper side. MCAR implies label-MAR, so by part (1): $\Delta \le 0$.

Lower side. We start from the chain–rule identity applied to $(A, B, C) = (\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}})$:

$$I(\mathbf{Y}; \mathbf{X}, \tilde{\mathbf{X}}) = I(\mathbf{Y}; \tilde{\mathbf{X}}) + I(\mathbf{Y}; \mathbf{X} \mid \tilde{\mathbf{X}}) = I(\mathbf{Y}; \mathbf{X}) + I(\mathbf{Y}; \tilde{\mathbf{X}} \mid \mathbf{X}).$$

Rearranging gives

$$-\Delta \;=\; I(\mathbf{Y}; \mathbf{X}) - I(\mathbf{Y}; \tilde{\mathbf{X}}) \;=\; I(\mathbf{Y}; \mathbf{X} \mid \tilde{\mathbf{X}}) - I(\mathbf{Y}; \tilde{\mathbf{X}} \mid \mathbf{X}). \tag{10}$$

The second term on the right is nonnegative, hence

$$-\Delta \;\le\; I(\mathbf{Y}; \mathbf{X} \mid \tilde{\mathbf{X}}). \tag{11}$$

Using the bound $I(U; V \mid W) \le H(V \mid W)$, we get

$$-\Delta \;\le\; H(\mathbf{X} \mid \tilde{\mathbf{X}}). \tag{12}$$

Index the matrix entries by a total order $\prec$ on pairs (i, j) and apply the chain rule:

$$H(\mathbf{X} \mid \tilde{\mathbf{X}}) = \sum_{(i,j)} H\big(X_{ij} \,\big|\, \tilde{\mathbf{X}},\, \{X_{kl} : (k, l) \prec (i, j)\}\big).$$

Since conditioning reduces entropy,

$$H(\mathbf{X} \mid \tilde{\mathbf{X}}) \;\le\; \sum_{i,j} H\big(X_{ij} \mid \tilde{X}_{ij}\big). \tag{13}$$

Fix (i, j) and denote $\pi_{ij} = \Pr[X_{ij} = 1]$. Under uniform MCAR,

$$\Pr[\tilde{X}_{ij} = ?] = \mu, \qquad \Pr[\tilde{X}_{ij} = x] = (1-\mu)\Pr[X_{ij} = x], \quad x \in \{0, 1\}.$$

Hence: (i) if $\tilde{X}_{ij} \in \{0, 1\}$ then X_{ij} is revealed, so $H(X_{ij} \mid \tilde{X}_{ij} \in \{0, 1\}) = 0$; (ii) if $\tilde{X}_{ij} = ?$, then $\Pr[X_{ij} = 1 \mid \tilde{X}_{ij} = ?] = \pi_{ij}$ and $H(X_{ij} \mid \tilde{X}_{ij} = ?) = h_2(\pi_{ij})$. Averaging over $\tilde{X}_{ij}$ gives

$$H(X_{ij} \mid \tilde{X}_{ij}) \;=\; \mu\, h_2(\pi_{ij}). \tag{14}$$

Combining (13) and (14):

$$H(\mathbf{X} \mid \tilde{\mathbf{X}}) \;\le\; \sum_{i,j} \mu\, h_2(\pi_{ij}) \;=\; nd\,\mu \cdot \frac{1}{nd} \sum_{i,j} h_2(\pi_{ij}) \;\le\; nd\,\mu \cdot h_2\left(\frac{1}{nd} \sum_{i,j} \pi_{ij}\right),$$

since h_2 is concave. Note that

$$\frac{1}{nd} \sum_{i,j} \pi_{ij} = \frac{1}{nd} \sum_{i,j} \Pr[X_{ij} = 1] = \mathbb{E}\left[\frac{1}{nd} \sum_{i,j} \mathbb{I}\{X_{ij} = 1\}\right] = 1 - \mathbb{E}[s(\mathbf{X})].$$

Using the symmetry $h_2(u) = h_2(1-u)$, we conclude

$$H(\mathbf{X} \mid \tilde{\mathbf{X}}) \;\le\; nd\,\mu \cdot h_2\big(\mathbb{E}[s(\mathbf{X})]\big).$$

Combining with $-\Delta \le H(\mathbf{X} \mid \tilde{\mathbf{X}})$ gives

$$-\,nd\,\mu\, h_2\big(\mathbb{E}[s(\mathbf{X})]\big) \;\le\; \Delta \;\le\; 0.$$

This concludes the proof. □

B Full Plots of Model Performance across Missingness

Figure 3 reports the full set of results (F1 score) across all proposed datasets (AIR, ELECTRIC, SYNTHETIC, TADPOLE) and missingness mechanisms, *Uniform Random MCAR* (UR), *Structural MCAR* (S) and *Correlation Driven MCAR* (CD). Each plot includes the ranking of models by their mean F1 score across all levels of missingness (μ). As can be seen, in almost all cases performance degrades much earlier and more severely than on Cora or Citeseer, confirming the higher difficulty and realism of the proposed datasets and mechanisms.

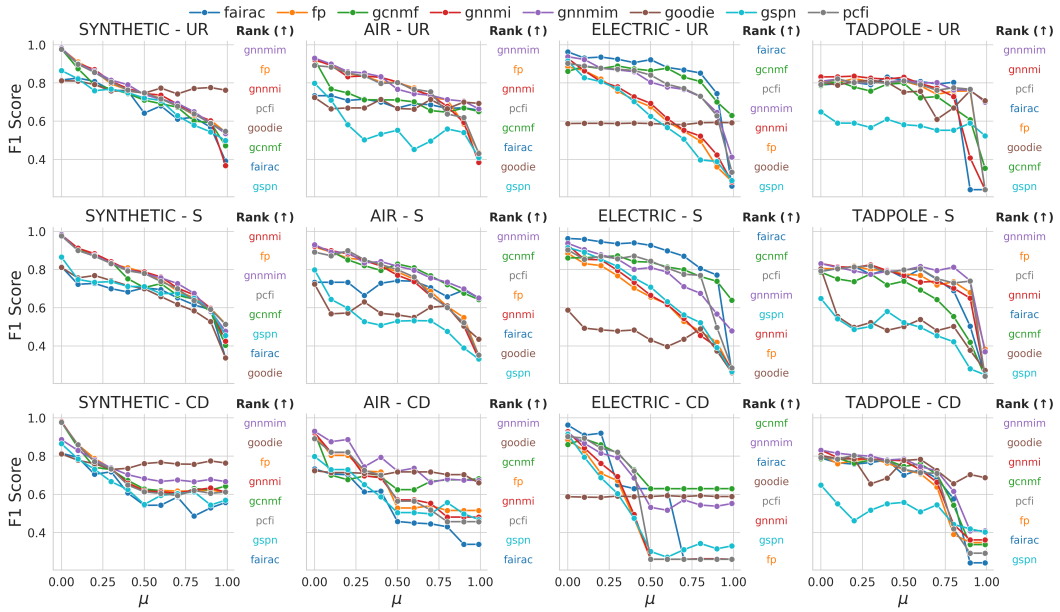

Figure 3: F1 score as a function of feature missingness (μ) for our proposed datasets (AIR, ELECTRIC, SYNTHETIC, TADPOLE), under different miss. mechanisms. To the right of each plot, models are sorted by their mean F1 rank across all μ values (best at the top)

C Experimental Details

All baseline and competitor methods are implemented using the official code released in their respective repositories, following the recommended training protocols and hyperparameter settings. For `GNNmi` and `GNNmim`, we adopt a standard GNN architecture where the convolutional layer type (chosen among GCN, GAT, GRAPHSAGE, and GIN), the number of layers (1-3), the learning rate (10^{-4}-10^{-2}), and the weight decay (10^{-5}-10^{-3}) are tuned via grid search on the validation set. All models are trained on the same data splits with early stopping to ensure a fair comparison.

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: The abstract and introduction state the contributions (datasets, GNNmim, evaluation settings) and these are supported by the theoretical and experimental results.

   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: The paper explicitly discusses the limitations of current benchmarks, the scope of the proposed datasets, and the evaluation settings, noting that larger benchmarks and additional missingness types remain open challenges.

   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: All theoretical results are stated with their assumptions, and complete proofs are provided in the appendix with references to the main text.

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: The paper details datasets, baselines, training protocols, and hyperparameter settings, ensuring that the main experimental results can be reproduced.

   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: [Yes]

Justification: Anonymized code and instructions, along with links to the datasets and preprocessing details, are provided in the supplemental material to enable faithful reproduction of the results.

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, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies datasets, splits, baselines, optimizers, and hyperparameters, with full details provided in the appendix and supplemental material.

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: Results are averaged over multiple runs with mean and standard deviation reported, and error bars are shown in figures to indicate variability across random seeds.

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: The paper specifies that experiments were run on GPU-equipped machines, with details on hardware and runtime provided in the appendix to enable reproduction.

   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: The work adheres to the NeurIPS Code of Ethics; it uses publicly available or synthetic datasets and raises no ethical concerns regarding data collection or deployment.

   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: The paper discusses positive impacts, such as enabling more reliable GNNs in domains like healthcare and sensor networks. No immediate negative societal impacts are anticipated.

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: [NA]

    Justification: Our results pose no such risks.

    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: We properly credit the original authors of all models and datasets used in our work, and we respect the licenses and terms of use associated with them

    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: The new assets include our synthetic dataset, and three real world datasets. We release well-documented code and data under a non-restrictive license

    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: The paper does not involve crowdsourcing experiments or research with human subjects

    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: The paper does not involve crowdsourcing experiments or research

    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: LLMs were not used as part of the core methods in this research.

    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.