
Energy-based Epistemic Uncertainty for Graph Neural Networks

Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann
School of Computation, Information and Technology & Munich Data Science Institute
Technical University of Munich, Germany
{d.fuchsgruber, tom.wollschlaeger, s.guennemann}@tum.de

Abstract

In domains with interdependent data, such as graphs, quantifying the epistemic uncertainty of a Graph Neural Network (GNN) is challenging as uncertainty can arise at different structural scales. Existing techniques neglect this issue or only distinguish between structure-aware and structure-agnostic uncertainty without combining them into a single measure. We propose GEBM, an energy-based model (EBM) that provides high-quality uncertainty estimates by aggregating energy at different structural levels that naturally arise from graph diffusion. In contrast to logit-based EBMs, we provably induce an integrable density in the data space by regularizing the energy function. We introduce an evidential interpretation of our EBM that significantly improves the predictive robustness of the GNN. Our framework is a simple and effective post hoc method applicable to any pre-trained GNN that is sensitive to various distribution shifts. It consistently achieves the best separation of in-distribution and out-of-distribution data on 6 out of 7 anomaly types while having the best average rank over shifts on *all* datasets.

1 Introduction

Quantifying and understanding uncertainty is crucial to developing safe and reliable machine learning systems. Many applications such as Reinforcement Learning [42], Active Learning [5, 31] or Out-of-Distribution detection [60] benefit from disentangling different facets of uncertainty [54, 17, 46]. Typically, one distinguishes between an irreducible *aleatoric* component and reducible *epistemic* factors [29]. The former is inherent to the data, while the latter is rooted in a lack of knowledge. One way to quantify epistemic uncertainty is to define a classifier-dependent density over the data domain [55]. High values indicate similarity to the training data and thus imply low epistemic uncertainty. Energy-based models (EBMs) induce this density by defining an energy function that assumes low values near the training data [38]. A common choice is the logits of the classifier as their magnitude is supposed to correlate with the model confidence [40]. However, many architectures have been shown to produce arbitrarily overconfident predictions far from the training data [28]. The negative implications for logit-based EBMs are only scarcely discussed in the literature [35, 36].

While substantial effort has been directed toward uncertainty estimation for independent and identically distributed (i.i.d.) problems [1, 24, 49, 37, 6, 22], graphs have only recently received attention within the community [71, 64, 67, 21]. There, uncertainty can arise at different structural scales, e.g., from only a single node, clusters, or global properties. Previous work only accounts for this implicitly by applying techniques from i.i.d. domains to Graph Neural Networks (GNNs) [67, 49]. A recent approach distinguishes only between structure-aware and structure-agnostic uncertainty [64] while not combining them into a single measure, which is often required in downstream applications [42, 54]. Consequently, estimates are often only sensitive to shifts that match their structural resolution. They may overfit certain anomaly types and not be reliable in general.

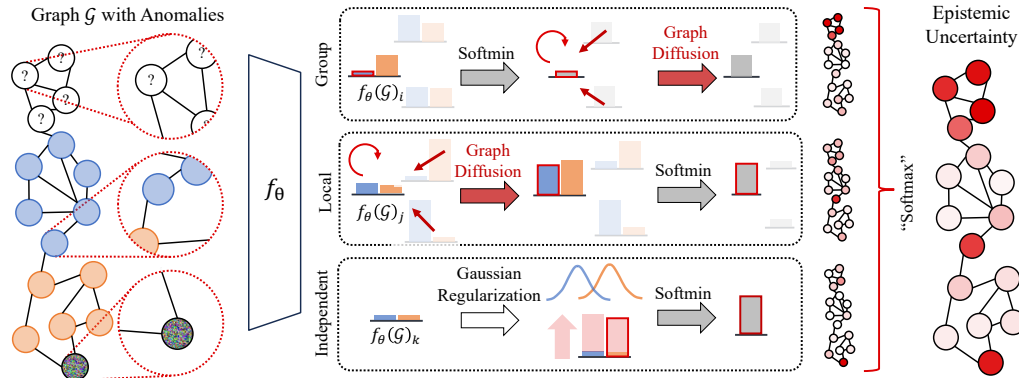


Figure 1: Overview of the Graph Energy-based Model (GEBM). Graph-agnostic energy (uncertainty) of a trained GNN $f_\theta(\mathcal{G})$ is first regularized to mitigate overconfidence and then aggregated at a local, cluster, and structure-independent scale by interleaving energy marginalization and graph diffusion. While group energy marginalizes before diffusion, local energy is fine-grained and can pick up conflicting evidence. GEBM assigns high uncertainty to several anomaly types simultaneously.

We address these shortcomings and propose a novel graph-based EBM (GEBM) that is structure-aware at different levels. The core idea behind GEBM is to define energy functions at different structural abstractions and facilitate the flexibility of the EBMs to aggregate them into a single, theoretically grounded measure. Interestingly, we find that interleaving a graph diffusion process with energy marginalization gives rise to energy functions that naturally capture patterns of different granularity. As depicted in Figure 1, we utilize three energy types: (i) *Group* energy corresponds to evidence smoothing and emphasizes anomalies clusters within the graph while not distinguishing between their type locally. (ii) *Local* energy is more granular and sensitive to evidence disagreements in the neighborhood of a node. (iii) *Independent* energy is fully structure-agnostic and describes patterns that are limited to individual nodes. Aggregating them using soft maximum selection induces a single energy measure that assigns high uncertainty to anomalies at various scales. A Gaussian regularizer provably ensures that GEBM converges to low confidence far from the training data.

We evaluate GEBM over an extensive suite of datasets, distribution shifts, and baselines¹. It consistently exhibits state-of-the-art performance in detecting out-of-distribution (o.o.d.) instances, while other approaches are only effective in a subset of settings. On all datasets, GEBM ranks the best on average over all distribution shifts. Beyond o.o.d. detection, we discover a novel theoretical connection between EBMs and evidential models. This interpretation of GEBM enables the GNN backbone to provide accurate predictions even under severe distribution shifts.

In summary, we tackle three main deficiencies of energy-based uncertainty for graphs by:

1. Proposing GEBM, an EBM that is aware of interdependence and provides a single uncertainty estimate incorporating different structural scales.
2. Formally showing that a Gaussian regularizer mitigates the overconfidence problem of logit-based EBMs and enables GEBM to provably induce an integrable data density.
3. Showing how to interpret GEBM as an evidential approach which considerably improves the predictive robustness of the classifier under distribution shifts.

2 Background

Semi-Supervised Node Classification. We consider (semi-supervised) node classification on an attributed graph $\mathcal{G} = (\mathcal{A}, \mathcal{X})$ with n nodes $\mathcal{V}$ and m edges $\mathcal{E}$ that is represented by an adjacency matrix $\mathcal{A} = \{0, 1\}^{n \times n}$ and a node feature matrix $\mathcal{X} \in \mathbb{R}^{n \times d}$. Each node has a label $y_i \in \{1, \dots, C\}$. Given the labels of training nodes, the task is to infer the labels of the remaining nodes. Lastly, we focus on homophilic graphs, i.e. edges are predominantly present between nodes of the same class.

¹We provide our code at cs.cit.tum.de/daml/gebm/

Uncertainty in Machine Learning. Usually, uncertainty regarding the prediction of a model is disentangled into aleatoric and epistemic factors u^{alea} and u^{epi} , respectively. The former encompasses irreducible sources of uncertainty like measurement noise and inherent ambiguities. The latter is commonly defined as the non-aleatoric components of the overall uncertainty and can in general, be reduced, for example by acquiring additional data.

Energy-based Models. Epistemic uncertainty is commonly defined to be anti-correlated to a classifier-dependent density $p_\theta(\mathbf{x})$ over the input space [55]. We study Energy-based Models (EBMs) [38] which define a joint energy $E_\theta(\mathbf{x}, y)$ over features and labels $p_\theta(\mathbf{x}, y) \propto \exp(-E_\theta(\mathbf{x}, y))$. Marginalization over the labels induces a feature density with normalizer Z_θ .

$$p_\theta(\mathbf{x}) = Z_\theta^{-1} \sum_y \exp(-E_\theta(\mathbf{x}, y)) = Z_\theta^{-1} \exp(-E_\theta(\mathbf{x})) \quad (1)$$

Intuitively, the energy $E_\theta(\mathbf{x}) = -\log \sum_y \exp(-E_\theta(\mathbf{x}, y))$ can be interpreted as a soft minimum of the joint energies. It is low in regions of high confidence and high elsewhere. The measure of epistemic uncertainty implied by an EBM therefore is: $u^{\text{epi}}(\mathbf{x}) = -\log p_\theta(\mathbf{x}) = E_\theta(\mathbf{x}) + \text{const}$. We can also write the conditional class probabilities as:

$$p_\theta(y | \mathbf{x}) = \exp(-E_\theta(\mathbf{x}, y)) / \sum_{y'} \exp(-E_\theta(\mathbf{x}, y')) \quad (2)$$

This is exactly the softmax distribution applied to $-E_\theta(\mathbf{x}, y)$ which directly connects the logits $f_\theta(\mathbf{x}) \in \mathbb{R}^C$ predicted by a classifier to the joint energy by defining $E_\theta(\mathbf{x}, y) := -f_\theta(\mathbf{x})_y$ [40].

Evidential Deep Learning. In contrast to first-order methods that directly predict $p(\mathbf{y} | \mathbf{x})$, evidential methods [62, 43] instead parameterize the corresponding second-order distribution from which first-order distributions are sampled [32]. In classification, this second-order distribution is a Dirichlet distribution with parameters $\alpha > \mathbf{0}$ called *evidence*. They indicate the confidence of the classifier. For inference, a first-order prediction can be obtained by taking the expectation:

$$\mathbb{E}_{\mathbf{p} \sim \text{Dir}(\alpha)} [p(\mathbf{y} | \mathbf{x})] = \alpha / \sum_{y'} \alpha_{y'} \quad (3)$$

3 Related Work

Uncertainty Estimation for i.i.d. Data. Disentangling uncertainty has been approached from various perspectives. A family of sampling-based approaches uses a Bayesian Information-theoretic framework [29] that relies on stochastic predictions derived from a posterior over model weights. The total uncertainty is defined on the mean predictor, epistemic uncertainty as the deviation of each sample from that mean, and aleatoric uncertainty as the difference of both estimates [1]. The posterior can either be explicitly modeled by Bayesian Neural Networks (BNNs) [6, 13, 15, 18], Monte-Carlo Dropout (MCD) [22], or implicitly realized by ensemble methods [37, 77]. Test-time augmentation [72, 73] and Stochastic Centering [66] also provide samples from this posterior. Sampling-free approaches are deterministic and estimate uncertainty with a single forward pass. Evidential methods [62, 68] predict a second-order distribution from which epistemic uncertainty is derived. Distance-based approaches quantify epistemic uncertainty as similarity to the training data [46, 48, 20, 65] and are closely related to density-based uncertainty estimation. (Deep) Gaussian Processes (GPs) [52, 45, 39] use a kernel function to measure this similarity but do not disentangle epistemic and aleatoric uncertainty. Posterior Networks combine evidential and density-based approaches by predicting density-based updates to the evidence [11, 10].

Uncertainty Estimation for Graphs. Many of these approaches transfer to graphs: DropEdge [59] applies MCD to edges and GPs can utilize a structure-aware kernel [53, 83, 41, 60]. SGCN [82] proposes an evidential student-teacher approach while G-ΔUQ [67] applies Stochastic Centering to a GNN and improves on calibration. Graph Posterior Network (GPN) [34] diffuses the density-based evidence of a Posterior Network but provides separate measures for structure-aware and structure-agnostic uncertainty, each of which is only effective on some distribution shifts.

Energy-based Models. EBMs [38, 3, 2] are typically employed in generative modelling [14] but have also been applied to uncertainty estimation [16] and anomaly detection [40, 81]. To address the overconfidence of logit-based EBMs far from training data, a Gaussian regularization term has

recently been proposed [35] that we employ in our framework as well. While their work empirically validates this adjustment, we formally prove the overconfidence of logit-based EBMs to happen with high probability and Gaussian regularization to mitigate the issue. The HEAT framework [36] learns multiple corrected EBMs via stochastic gradient Langevin dynamics [76] and composes them into a single measure. In contrast, our approach does not require additional training and aggregates energy that emerges naturally at different scales in the graph from a single logit-based joint energy model. Lastly, GNNSafe [79] diffuses the logit-based energy of a GNN to improve its out-of-distribution detection. In contrast, our model considers energy beyond the cluster scale.

4 Method

We develop a simple yet effective **Graph-Energy-Based Model (GEBM)** for post hoc epistemic uncertainty that is sensitive to a variety of distribution shifts from any pre-trained GNN.

4.1 Energy at Different Scales

Previous work either does not distinguish between uncertainty at different structural resolutions at all [79] or only disentangles structure-aware and structure-agnostic factors without combining them [64]. However, for many practical purposes, epistemic uncertainty must be quantified with a single measure [42, 5, 60]. We address these issues by proposing an EBM-based uncertainty that incorporates patterns on different natural abstractions of a graph. The density induced by GEBM, $p_\theta(\mathbf{x})$, serves as a singular uncertainty measure that is sensitive to multiple distribution shifts simultaneously.

Inspired by the success of graph diffusion [9, 23], we propose energy on different structural levels: (i) Graph-agnostic, for node features in isolation. (ii) Based on the evidence in the local neighborhood of a node. (iii) Within clusters in the graph. Note that defining global energy on the whole graph induces no differences at the node level and therefore we do not consider it in this work. We point to appropriate measures for a potential extension of GEBM in the existing literature [84].

We make the intriguing observation that interleaving a diffusion process $P_A : \mathbb{R}^k \rightarrow \mathbb{R}^k$ with the marginalization of structure-agnostic joint energy $E_\theta(\mathbf{x}, y)$ as in Equation (1) induces energy functions that describe the aforementioned natural abstraction levels. Based on this insight, we can derive definitions for three types of energy functions:

Independent Energy. On the finest scale, we consider energy independent of structural effects by omitting the diffusion operator. This term captures uncertainty regarding node features in isolation.

$$E_{\theta,I}(\mathbf{x}) = -\log \sum_y \exp(-E_\theta(\mathbf{x}, y)) \quad (4)$$

Local Energy. On a coarser scale, we diffuse the joint energy *before* marginalization. This retains local information like conflicting feature-based evidence within the neighborhood of a node as the class-specific information is marginalized *after* propagation:

$$E_{\theta,L}(\mathbf{x}) = -\log \sum_y \exp[P_A(-E_\theta(\mathbf{x}, y))] \quad (5)$$

Group Energy. By interchanging marginalization and diffusion, we effectively smooth the marginal evidence $E_\theta(\mathbf{x})$. Therefore, energy gets propagated predominantly within clusters of the graph.

$$E_{\theta,G}(\mathbf{x}) = -P_A \left[\log \sum_y \exp(-E_\theta(\mathbf{x}, y)) \right] \quad (6)$$

Since marginalization is done *before* propagation, less local information is preserved: The energy of a node will increase when its cluster is anomalous, regardless of whether the type of anomaly matches its own. As can be seen exemplarily in Figure 1, this loss of local information comes at the benefit of being less exposed to local variability within coarser clustered patterns.

Each energy type captures anomalies that affect the corresponding structural scale. We provide synthetic experiments in Appendix C.7 as an additional intuitive explanation for the aforementioned energy terms. In practice, GEBM enables practitioners to augment our framework with further, potentially task-specific, energy functions. We empirically find however that the combination of the three naturally arising energy terms already detects a broad range of distribution shifts.

4.2 Regularizing Logit-based EBMs

Following Equation (1), EBMs imply a density over the data domain $p_\theta(\mathbf{x})$ with normalization constant:

$$Z_\theta = \int_{\mathcal{X}} \sum_y \exp(-E_\theta(\mathbf{x}, y)) d\mathbf{x} \quad (7)$$

When quantifying epistemic uncertainty with this density, the energy must be low in regions of the input feature space close to the training data, while high values should be assumed far away. To that end, $p_\theta(\mathbf{x})$ must be integrable, that is, have a finite normalizer Z_θ . However, previous work has shown that piecewise affine classifiers which are the backbone of modern GNNs will converge to overconfident logits far from training data [28]. We remark that previous studies found GNNs to be underconfident on in-distribution data [74, 75] while the aforementioned issue of overconfidence arises from high distance to training data induced by a distribution shift (see Appendix C.8). This overconfidence is problematic, as logit-based energy can suffer from the same issue and $E_\theta(\mathbf{x})$ may assume arbitrarily small values far from the training data. This contradicts the aforementioned desideratum of low confidence and the implied uncertainty measure breaks under severe distribution shifts. The normalizer of the logit-based joint energy Z_θ is not finite with a high probability and therefore the EBM can not induce an integrable density $p_\theta(\mathbf{x})$.

Proposition 4.1. *Let $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^C$ be a piecewise affine function and $\mathbb{R}^h = \bigcup_l^L Q_l$ be the disjoint set of polytopes on which f_θ is affine, i.e. $f_\theta(\mathbf{x}) = \mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)}$ for $\mathbf{x} \in Q_l$. Assuming the direction of the rows of each $\mathbf{W}^{(l)}$ to be uniformly distributed, the probability that Z_θ converges decreases exponentially in the number of non-closed linear regions L' and classes C .*

$$\Pr[Z_\theta < \infty] = (1/2)^{C \cdot L'}$$

We provide proofs for all claims in Appendix A. Intuitively, since f_θ behaves like an affine function in the limit $\|\mathbf{x}\|_2 \rightarrow \infty$, its predicted logits may diverge toward ∞ . If for any class the model produces overconfident predictions in one of its affine regions, the marginal energy $E_\theta(\mathbf{x})$ will diverge toward maximal confidence. As classifiers are trained to output high values for one of the classes, we expect the actual probability of a well-behaved energy function to be even lower than our theoretical bound that assumes uniform weight directions. In practice, this pathological behavior of logit-based EBMs has been observed in previous work [64, 36, 35]. We mitigate this issue by augmenting the logit-based energy and we prove that this regularization induces an integrable density.

4.3 Our Model

Similarly to recent work on EBMs [35], we employ a class-conditional Gaussian prior $\mathcal{N}(\mathbf{f}_\theta(\mathbf{x})_y | \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ as a regularizer for the logit-based energy. We learn the parameters $\{\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\}_{y=1}^C$ as the maximum likelihood estimates from the training instances of each class.

$$\hat{E}_\theta(\mathbf{x}, y) = E_\theta(\mathbf{x}, y) - \gamma \log \mathcal{N}(\mathbf{f}_\theta(\mathbf{x})_y | \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \quad (8)$$

The regularization strength $\gamma > 0$ and the choice of the diffusion operator $P_{\mathbf{A}}$ are the only hyperparameters of GEBM. Each of them has an intuitive interpretation: γ controls the trade-off between the predictive dependency and distance-awareness of the EBM while $P_{\mathbf{A}}$ encapsulates how information is propagated over $\mathcal{G}$. Regularization ensures that the corresponding marginal density $\hat{p}_\theta(\mathbf{x})$ has a finite normalizer $\hat{Z}_\theta$ and is therefore integrable.

Theorem 4.2. *For a piecewise affine classifier f_θ as in Proposition 4.1, $\hat{p}_\theta$ is well-defined.*

$$\hat{Z}_\theta = \int_{\mathcal{X}} \sum_y \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(\mathbf{f}_\theta(\mathbf{x})_y | \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)^\gamma d\mathbf{x} < \infty$$

This regularized joint energy diverges toward maximal uncertainty in the limit. Consequently, its induced density provides an uncertainty estimator that is reliable even far away from the training data.

Corollary 4.3. *For a piecewise affine classifier f_θ as in Proposition 4.1, and any $\mathbf{x} \in \mathbb{R}^d$ almost surely:*

$$\lim_{\alpha \rightarrow \infty} \hat{p}_\theta(\alpha \mathbf{x}) = 0$$

We then combine the *regularized energy* $\hat{E}_\theta(\mathbf{x}, y)$ at different structural scales (Equations (4) to (6)).

$$\hat{E}_{\theta, \text{GEBM}}(\mathbf{x}) = \log \left[\exp \left(\hat{E}_{\theta, I}(\mathbf{x}) \right) + \exp \left(\hat{E}_{\theta, L}(\mathbf{x}) \right) + \exp \left(\hat{E}_{\theta, G}(\mathbf{x}) \right) \right] \quad (9)$$

Since energy is defined per node, most diffusion processes (Appendix C.6) act on individual nodes as an affine function with a positive coefficient (see Appendix A.2). Therefore, each individual energy and their aggregate, $\hat{E}_{\theta, \text{GEBM}}$, induce an integrable density when using a linear diffusion operator $P_{\mathbf{A}}$.

Theorem 4.4. *For a linear diffusion operator $P_{\mathbf{A}}(\mathbf{x}) = \alpha \mathbf{x} + \text{const}$, $\alpha > 0$ and the regularized energy $\hat{E}_\theta(\mathbf{x}, y)$, GEBM induces an integrable density:*

$$\int_{\mathcal{X}} \exp \left(-\hat{E}_{\theta, \text{GEBM}}(\mathbf{x}) \right) d\mathbf{x} < \infty$$

GEBM is lightweight and can be applied post hoc to *any* logit-based GNN without additional training. Its induced density is a proxy for epistemic uncertainty: $u^{\text{epi}}(\mathbf{x}) = -\log \hat{p}_\theta(\mathbf{x})$ and we compute aleatoric uncertainty as the entropy of the predictive distribution of the GNN [29]. Following [79], we realize the diffusion operator as a repeated application of a label-propagation scheme (see Appendix C.6). Intuitively, this operator is a smoothing process which is only appropriate when assuming the in-distribution data to be homophilic. In the case of non-smooth (heterophilic) training data, the proposed energy would be high for in-distribution data which is undesired behaviour for an EBM-based uncertainty estimator. To disentangle effects at different structural scales, we define each of GEBM's components on structure-agnostic regularized joint energies $\hat{E}_\theta(\mathbf{x}, y)$. To that end, we compute the outputs of the classifier by omitting the structure *for the computation of $\hat{E}_\theta$ only* by evaluating $f_\theta(\mathbf{I}, \mathbf{X})$, i.e. setting $\mathbf{A} = \mathbf{I}$. These outputs depend only on the node features.

4.4 EBMs as Evidential Models

Beyond using epistemic uncertainty to detect anomalies, we also show how the joint density induced by GEBM enables predictions that are robust against distribution shifts. First, we discover a correspondence between first-order predictions of logit-based classifiers (Equation (2)) and evidential predictions (Equation (3)). This connects the joint energy $E_\theta(\mathbf{x}, \mathbf{y})$ of an EBM to the evidence α :

$$\alpha \simeq \exp(-E_\theta(\mathbf{x}, \mathbf{y})) \propto p_\theta(\mathbf{x}, \mathbf{y}) \quad (10)$$

For an integrable density $p_\theta(\mathbf{x}, \mathbf{y})$, the evidence α will vanish in the limit far away from training data. Previous work on Posterior Networks [11, 10] fits a normalizing flow to obtain class-conditional densities $p_\theta(\mathbf{x} \mid \mathbf{y})$ and uses them to compute a Bayesian update to a prior evidence α^{prior} . Its extension to graphs, GPN [64], diffuses these structure-agnostic updates with an operator $P_{\mathbf{A}}$. Similarly, our GEBM framework induces a normalized joint density $\hat{p}_\theta(\mathbf{x}, \mathbf{y})$ through its regularized energy function at no additional cost. Therefore, it can also be interpreted as an evidential classifier that enables inference according to Equation (3).

$$\overbrace{\alpha^{\text{post}} = \alpha^{\text{prior}} + P_{\mathbf{A}}(N * p_\theta(\mathbf{x}, \mathbf{y}))}^{\text{GPN [64] (Evidential)}} \quad \overbrace{\alpha^{\text{post}} = \alpha^{\text{prior}} + P_{\mathbf{A}}\left(Z_\theta^{-1} * \exp(-\hat{E}_\theta(\mathbf{x}, \mathbf{y}))\right)}^{\text{GEBM (ours)}}$$

Here, N is called an uncertainty budget and it roughly corresponds to the normalizer of the joint energy. This interpretation of EBMs recovers desirable properties of density-based evidential methods [11, 64]: Predictions will converge toward a prior α^{prior} far from the training distribution. They will be less affected by anomalies in the neighborhood of a node and therefore be more robust against distribution shifts. We remark that while our framework enables this evidential inference scheme, it is also possible to instead use the backbone classifier as-is and preserve its predictive performance.

5 Experiments

We evaluate the efficacy of GEBM by extending the evaluation proposed in [64] and expose it to a suite of 7 distribution shifts from three families that cover a broad range of anomaly types.

5.1 Setup

Datasets and Distribution Shifts. We use seven common benchmark datasets for node classification: The five citation datasets *CoraML* [4], *CoraML-LLM* [4, 78] [4], *Citeseer* [61, 25], *PubMed* [51], *Coauthor-Physics* and *Coauthor-Computers* [63], and the co-purchase graphs *Amazon Photos* and *Amazon Computers* [44]. We expose all methods to three families of distribution shifts:

- (i) **Structural.** We select nodes with the lowest *homophily* as o.o.d. and train on homophilic nodes only. This induces a shift in the local connectivity pattern of the nodes. Furthermore, we include a setting in which nodes with low *Page-Rank* (PR) centrality [56] are considered o.o.d.
- (ii) **Leave-out-Class.** We withhold nodes belonging to a subset of classes during training and reintroduce them during inference. In practice, this might, for example, correspond to a new type of user joining a social network. We either select the held-out classes randomly or pick those with the lowest average homophily to make them more dissimilar to the retained classes.
- (iii) **Feature Perturbations.** We randomly choose a subset of nodes and perturb their features by replacing them with noise. Since most datasets have categorical bag-of-words features, we have fine-grained control over the severity of the shift in this setting. We generate *near-o.o.d.* data by sampling from a Bernoulli distribution $\bar{p}$ fitted on the node features of the dataset $\mathbf{X}$. A stronger shift is induced by fixing the success probability at a value of $p = 0.5$. Drawing node features from $\mathcal{N}(0, 1)$ constitutes a domain shift (*far-o.o.d.*) for categorical data.

Existing work concerns transductive node classification which enables leakage of o.o.d. information during training [64, 79]. Instead, we study the inductive setting and remove o.o.d. nodes and their edges during training, and provide results for the transductive setting in Appendix C. All results were averaged over 5 splits and 5 initializations each (for standard deviations, see Appendix C).

Model and Baselines. We compare GEBM to different baselines for (epistemic) uncertainty estimation: Ensembles (**GCN-Ens**), MC-dropout (**GCN-MCD**), DropEdge (**GCN-DropEdge**), a combination of MC-dropout and DropEdge (**GCN-MCD+DropEdge**) and a Bayesian GCN (**BGCN**) are sampling-based approaches. We also compare against the logit-based EBM (**GCN-Energy**), **GNNSafe**, and **HEAT** as EBM baselines. Lastly, we consider **GPN** and **SGCN** as evidential methods. For all models, including GEBM, we use the same backbone architecture (see Appendix B.2).

Metrics. We evaluate epistemic uncertainty by detecting distribution shifts. That is, we report the AUC-ROC and AUC-PR metrics for the binary *out-of-distribution detection* problem of separating in-distribution (i.d.) from out-of-distribution (o.o.d.) nodes. Additionally, we report how well uncertainty correlates with erroneous (*misclassification detection*) in Appendix C.3. Since our proposed method does not alter the softmax predictions of the backbone model, it will affect neither the aleatoric estimates nor their *calibration*. Consequently, our framework is open to additional post hoc calibration methods such as temperature scaling [26]. For completeness, we report the Expected Calibration Error (ECE) [50] and the Brier score [7] in Appendix C.4 for the GNN backbone.

5.2 Results

Out-of-Distribution Detection. Table 1 shows the performance of different aleatoric and epistemic uncertainty methods on various distribution shifts (full results in Table 6). Across all datasets and distribution shifts, our model provides the best or second-best separation of o.o.d. data from i.d. data. In practice, an estimator that is effective on all distribution shifts simultaneously is desirable. Therefore, we rank ($\downarrow$) all estimators individually for each shift and dataset and report its average rank over all distribution shifts for each dataset. To not favor one distribution shift family, we adjust the weight such that Leave-out-Class, structural shifts, and feature perturbations contribute the same amount. We rank all epistemic uncertainty proxies both against other epistemic measures and aleatoric uncertainty separately. In both cases, Tables 2 and 7 show that our proposed EBM-based epistemic uncertainty is the only estimator that is effective under different distribution shifts at the same time. On all datasets, GEBM improves the AUC-ROC scores by 5.8 to 10.9 percentage points on average (16%-32% relative improvement) over a vanilla EBM, the second best-ranked estimator (Appendix C.2). Since all EBM-based approaches (GCN-EBM, GCNSafe, GEBM) are post hoc methods, they share the aleatoric uncertainty and high predictive accuracy of the backbone.

Out of all 7 distribution shifts, our framework is only less sensitive to the centrality shift. Many baselines only perform well because of their symmetric diffusion operator that biases any arbitrary

	Model	LoC (<i>last</i>)		Ber(0.5)		$\mathcal{N}(0, 1)$ (<i>far</i>)		Homophily	
		AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$
		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)	
CoraML	GCN-DE	86.2/73.0	87.2	47.1/ 68.8	71.9	26.8/57.8	71.9	70.8/54.5	85.6
	GCN-Ens	89.8/73.9	90.4	59.4/64.1	75.9	30.4/52.7	75.9	71.5/68.5	91.5
	GPN	85.3/88.1	88.5	53.4/52.4	72.8	51.4/55.5	72.8	66.9/51.5	87.6
	GCN-EBM	89.7/ 89.9	90.3	59.9/58.6	75.8	31.7/25.7	75.8	71.3/71.2	91.2
	GCN-HEAT	89.7/87.1	90.3	59.9/64.4	75.8	31.7/ 76.0	75.8	71.3/70.0	91.2
	GCNSafe	89.7/ 91.6	90.3	59.9/53.4	75.8	31.7/35.3	75.8	71.3/ 73.1	91.2
	GCN-GEBM	89.7/ 91.6	90.3	59.9/ 94.5	75.8	31.7/ 86.4	75.8	71.3/ 76.7	91.2
Amazon Photo	GCN-DE	76.4/61.2	91.7	52.8/51.4	88.5	41.8/51.5	88.5	62.7/52.0	96.6
	GCN-Ens	74.7/80.3	91.9	52.3/51.3	89.9	42.6/55.0	89.9	62.0/65.2	97.5
	GPN	74.5/ 87.3	88.8	55.3 /54.9	86.5	54.2/55.5	86.5	65.9 /47.0	97.1
	GCN-EBM	74.6/75.2	91.8	52.1/52.1	89.8	44.2/42.6	89.8	61.9/60.0	97.4
	GCN-HEAT	74.6/77.3	91.8	52.1/51.8	89.8	44.2/ 57.3	89.8	61.9/62.6	97.4
	GCNSafe	74.6/75.6	91.8	52.1/50.9	89.8	44.2/46.4	89.8	61.9/60.5	97.4
	GCN-GEBM	74.6/ 83.6	91.8	52.1/ 64.2	89.8	44.2/ 92.4	89.8	61.9/ 68.3	97.4
Coauthor-CS	GCN-DE	89.2/73.7	91.3	64.3/67.9	88.8	33.7/60.3	88.8	62.9/55.7	97.1
	GCN-Ens	89.7/87.3	92.4	69.6/67.8	90.0	32.8/60.7	90.0	63.9/ 69.1	97.9
	GPN	73.7/88.3	86.6	54.1/59.7	81.7	53.6/59.9	81.7	61.3/45.5	97.9
	GCN-EBM	89.7/89.9	92.3	69.3/ 71.6	89.8	34.7/28.8	89.8	63.8/60.5	97.8
	GCN-HEAT	89.7/88.6	92.3	69.3/70.0	89.8	34.7/ 69.8	89.8	63.8/66.6	97.8
	GCNSafe	89.7/ 92.1	92.3	69.3/63.0	89.8	34.7/36.9	89.8	63.8/62.4	97.8
	GCN-GEBM	89.7/ 92.8	92.3	69.3/ 99.5	89.8	34.7/ 90.0	89.8	63.8/ 69.5	97.8

Table 1: Out-of-Distribution detection AUC-ROC ($\uparrow$) using aleatoric or epistemic uncertainty (**best** and **runner-up**). Our epistemic measure achieves the strongest performance on most datasets and shifts while maintaining the classification accuracy of the GCN backbone.

Model	CoraML	CoraML LLM	Citeseer	PubMed	Amazon Computers	Amazon Photo	Coauthor CS	Coauthor Physics
GCN-DE	6.3/14.2	8.0/16.9	7.8/16.2	7.3/13.6	7.2/15.9	7.1/15.8	6.5/13.2	8.1/16.8
GCN-Ens	7.8/16.1	6.3/15.0	6.4/12.9	6.2/12.4	5.7/11.6	4.7/ 9.0	4.8/10.4	6.2/13.3
GPN	7.0/15.2	5.4/11.2	6.6/12.9	5.8/12.4	5.7/11.8	5.3/11.7	7.2/13.9	6.1/14.5
GCN-EBM	4.5/ 7.9	4.0 /8.2	5.2/10.3	5.2 /12.0	4.2/10.7	4.9/10.9	4.1 / 7.1	4.4 / 8.4
GCN-HEAT	4.4 /10.4	5.4/12.9	3.8 / 9.2	5.6/12.8	3.5 / 8.6	4.5 /10.2	4.2/8.3	4.7/10.4
GCNSafe	5.3/9.1	4.4/ 8.1	5.2/9.5	5.4/ 8.9	5.9/13.2	6.4/12.9	5.5/10.7	5.8/11.7
GCN-GEBM	2.7 / 4.5	2.9 / 4.9	3.3 / 5.7	3.8 / 7.4	2.5 / 4.2	2.7 / 4.5	2.7 / 4.3	3.0 / 5.0

Table 2: Average o.o.d. detection rank ($\downarrow$) of epistemic uncertainty versus other epistemic measures / all uncertainty measures over all distribution shifts (**best** and **runner-up**). GEBM has the best average performance rank over all distribution shifts and epistemic (and aleatoric) uncertainty estimators.

signal toward high-degree nodes regardless of the quality of the uncertainty (see Appendix C.6). While this bias could be explicitly incorporated into GEBM by defining an additional energy term, the goal of our work is not to engineer our measure toward certain downstream settings. GEBM is already sensitive to many distribution shifts as-is. Even after accounting for this setting, our estimator achieves the best average rank over all shifts. We defer further discussion to Appendix B.

Misclassification Detection. In alignment with previous work [64], we observe aleatoric uncertainty to be more effective for misclassification detection than epistemic estimates. Our method performs competitively among other epistemic measures as shown in Table 16. As previously discussed, improving aleatoric uncertainty for GNNs is beyond the scope of this work.

Robust Evidential Inference. As described in Section 4.4, GEBM enables evidential inference. We expose our model to feature shifts sampled from $\mathcal{N}(0, 1)$ and increase the fraction of perturbed nodes. Figure 2 shows the predictive performance of the classifier induced by an evidential interpretation of the EBM at different regularization weights γ compared to a vanilla EBM and the evidential model, GPN. We maintain high accuracy like an evidential model, while the performance of the baselines rapidly deteriorates. The hyperparameter γ can be seen as interpolating between an unregularized overconfident logit-based EBM and an uncertainty-aware evidential method. While GPN requires explicit evidential training, our method enables robust inference at negligible additional cost from a pre-trained GNN. Furthermore, GPN notably sacrifices predictive capability on clean data, while GEBM enables more control over the trade-off between accuracy on clean and perturbed data.

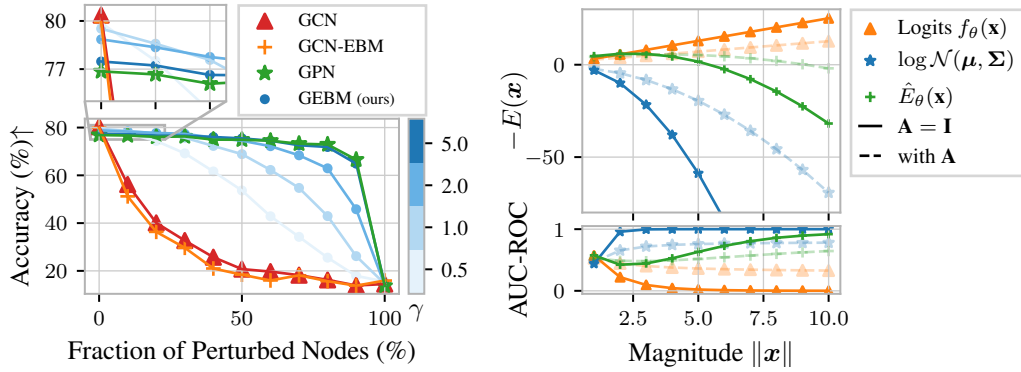


Figure 2: Accuracy of evidential inference at different regularization γ . GEBM performs on-par with evidential methods at increasingly severe perturbations.

Figure 3: Logit-based, Gaussian, and regularized energy at different magnitudes. Logit-EBMs become overconfident while the corrected energy is eventually dominated by its regularizer.

The advantage of this evidential perspective is that the evidence approaches zero under increasingly severe distribution shifts. Therefore, a fixed number of diffusion steps suffices to effectively counteract the influence of anomalous neighbors when making predictions for a node. In contrast, using diffusion at, for example, the predictive level is inadequate: As shown in Proposition 4.1, the energy is likely to diverge and, therefore, the number of diffusion steps necessary to mitigate arbitrarily severe distribution shifts is unbounded. Consequentially, we find that even compared to models like APPNP that heavily rely on graph diffusion at the predictive level, the evidential interpretation of GEBM notably improves robustness (see Figure 6).

5.3 Ablations.

Energy at Different Scales. In Table 3, we compare GEBM to energy at specific structural scales and a variant without regularization ($\gamma = 0$). Each of the former corresponds to one of the energies that the GEBM framework is composed of. Group energy is equivalent to a regularized logit-based EBM which is therefore also ablated implicitly. As expected, each component is sensitive to different shifts. This confirms that interleaving marginalization and diffusion captures patterns at different resolutions. As shown in Table 26, a variant of the aggregate GEBM achieves the best rank on 7 of 8 datasets. This shows that scale-awareness is the key ingredient for effective uncertainty estimation on graphs which can be further improved with energy regularization. In far-o.o.d. settings, this regularization is crucial to obtain reliable estimates as it mitigates overconfidence issues.

Model	LoC	$\mathcal{N}(0, 1)$	Homo.	Rank($\downarrow$)
EBM	87.5	24.7	67.9	4.1
Indep.	79.2	99.7	63.9	3.7
Local	86.9	62.1	79.8	4.3
Group	91.7	62.8	68.9	3.3
GEBM- E_θ	90.1	1.4	71.1	3.1
GEBM	89.7	89.8	73.1	2.6

Table 3: O.o.d.-detection AUC-ROC($\uparrow$) using different EBMs (best, runner-up). GEBM combines the benefits of energy at different scales and ranks best over splits.

Energy Regularization. We also ablate the effect of regularizing the joint logit-based energy in Figure 3. At increasing distance from the training data, logit-based energy becomes overconfident. In contrast, the regularized E_θ follows the logit-based energy near the training data and is dominated by the regularizer far away. This is reflected in the clear separation between perturbed and unperturbed nodes. When computing structure-aware energy, feature corruptions affect unperturbed nodes through the diffusion operation, making a clean separation difficult. This justifies our choice to compute structure-agnostic joint energy $\hat{E}_\theta(x, y)$ and factor in the graph afterward by applying P_A .

Backbone Architecture. Our method can be applied post hoc to any logit-based GNN. Table 4 evaluates the o.o.d.-detection performance of our EBM framework using differ-

	Model	LoC	Ber($\hat{p}$)	$\mathcal{N}(0, 1)$	Homo.
GCN	EBM	89.9	67.1	26.4	71.2
	Safe	91.6	56.9	36.1	73.1
	GEBM	91.6	77.1	86.6	76.7
GAT	EBM	90.2	55.6	44.1	72.1
	Safe	91.5	53.2	45.4	70.3
	GEBM	85.0	69.3	76.5	72.6
GIN	EBM	76.5	52.3	42.8	53.2
	Safe	79.0	49.3	46.2	51.2
	GEBM	80.7	51.4	53.6	65.4
SAGE	EBM	74.0	52.7	42.2	53.2
	Safe	77.3	49.2	46.6	51.1
	GEBM	77.3	51.6	54.7	62.7

Table 4: O.o.d. detection AUC-ROC($\uparrow$) using different backbones. Our method is effective on all architectures.

ent commonly used GNN backbones: GCN [34], GAT (v2) [70, 8], GIN [80] and GraphSAGE [27]. Standard deviations and AUC-PR are reported in Appendix C.5. Over different distribution shifts, our proposed approach consistently outperforms the logit-based EBM and the graph-specific GNNSafe variation. This shows the broad applicability of our framework that enables reliable high-quality epistemic uncertainty estimation from a large family of logit-based GNNs.

6 Limitations and Broader Impact

Limitations. As GEBM is a post hoc epistemic estimator, it does not improve aleatoric uncertainty or its calibration. In particular, the GCN backbone used in this work does not consistently achieve the strongest performance in both tasks. While the structural scales that arise naturally from graph diffusion cover many distribution shifts as-is, specific applications may require augmenting GEBM with additional energy terms, as we also observe for centrality-based shifts. While GEBM enables robust evidential inference, future work may build upon its paradigm of aggregating different structural scales in the graph for fully evidential methods. Lastly, we study homophilic node classification problems and leave an extension of GEBM beyond this setting to future work.

Broader Impact. GEBM enables cheap and simple uncertainty quantification for GNNs which we believe to contribute to the development of more reliable AI. Nonetheless, we encourage practitioners to actively reevaluate our measure of uncertainty to mitigate risks in safety-critical domains.

7 Conclusion

We propose GEBM, a simple and efficient EBM for post-hoc epistemic uncertainty estimation for GNNs. To the best of our knowledge, we are the first to consider uncertainty at different structural scales and address this gap by proposing a model that aggregates energy that naturally arises from graph diffusion. It consistently outperforms existing approaches over an extensive suite of datasets at detecting various distribution shifts. We formally and empirically confirm that logit-based EBMs suffer from overconfidence and prove that the regularized GEBM mitigates this issue by inducing an integrable data density. We exploit this property by discovering a link to evidential methods that enables the backbone to provide accurate predictions even under severe distribution shifts.

Acknowledgements

We want to give special thank to Leo Schwinn and Franz Rieger for giving helpful suggestions on an early draft of the manuscript. The research presented has been performed in the frame of the RADELN project funded by TUM Georg Nemetschek Institute Artificial Intelligence for the Built World (GNI). It is further supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy with funds from the Hightech Agenda Bayern.

References

- [1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.
- [2] David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169, 1985.
- [3] Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. *arXiv preprint arXiv:2003.05033*, 2020.
- [4] Sanghamitra Bandyopadhyay, Ujjwal Maulik, Lawrence B Holder, Diane J Cook, and Lise Getoor. Link-based classification. *Advanced methods for knowledge discovery from complex data*, pages 189–207, 2005.
- [5] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

- [6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks, 2015.
- [7] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.
- [8] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.
- [9] Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International Conference on Machine Learning*, pages 1407–1418. PMLR, 2021.
- [10] Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. *arXiv preprint arXiv:2105.04471*, 2021.
- [11] Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020.
- [12] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- [13] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Learning and policy search in stochastic dynamical systems with bayesian neural networks. *arXiv preprint arXiv:1605.07127*, 2016.
- [14] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [15] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792. PMLR, 2020.
- [16] Sven Elflein, Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. On out-of-distribution detection with energy-based models, 2021.
- [17] Derek Everett, Andre T Nguyen, Luke E Richards, and Edward Raff. Improving out-of-distribution detection via epistemic uncertainty adversarial training. *arXiv preprint arXiv:2209.03148*, 2022.
- [18] Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. *Advances in Neural Information Processing Systems*, 33:4346–4357, 2020.
- [19] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [20] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.
- [21] Dominik Fuchsgruber, Tom Wollschläger, Bertrand Charpentier, Antonio Oroz, and Stephan Günnemann. Uncertainty for active learning on graphs, 2024.
- [22] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.
- [23] Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. *Advances in neural information processing systems*, 32, 2019.

- [24] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.
- [25] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. Citeseer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98, 1998.
- [26] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.
- [27] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- [28] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019.
- [29] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.
- [30] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079, 2019.
- [31] Ajay J. Joshi, Fatih Murat Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379, 2009.
- [32] Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods?, 2024.
- [33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [34] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [35] Marc Lafon, Clément Rambour, and Nicolas Thome. Energy correction model in the feature space for out-of-distribution detection. *arXiv preprint arXiv:2403.10403*, 2024.
- [36] Marc Lafon, Elias Ramzi, Clément Rambour, and Nicolas Thome. Hybrid energy based model in the feature space for out-of-distribution detection, 2023.
- [37] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.
- [38] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fugie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- [39] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.
- [40] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2021.
- [41] Zhao-Yang Liu, Shao-Yuan Li, Songcan Chen, Yao Hu, and Sheng-Jun Huang. Uncertainty aware graph gaussian process for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4957–4964, 2020.

- [42] Owen Lockwood and Mei Si. A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pages 155–162, 2022.
- [43] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [44] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.
- [45] Pablo Morales-Alvarez, Daniel Hernández-Lobato, Rafael Molina, and José Miguel Hernández-Lobato. Activation-level uncertainty in deep neural networks. In *International Conference on Learning Representations*, 2020.
- [46] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv preprint arXiv:2102.11582*, 2021.
- [47] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394, June 2023.
- [48] Jishnu Mukhoti, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.
- [49] Sai Munikoti, Deepesh Agarwal, Laya Das, and Balasubramaniam Natarajan. A general framework for quantifying aleatoric and epistemic uncertainty in graph neural networks. *Neuro-computing*, 521:1–10, 2023.
- [50] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.
- [51] Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, volume 8, page 1, 2012.
- [52] Yin Cheng Ng, Nicolò Colombo, and Ricardo Silva. Bayesian semi-supervised learning with graph gaussian processes. *Advances in Neural Information Processing Systems*, 31, 2018.
- [53] Yin Cheng Ng, Nicolò Colombo, and Ricardo Silva. Bayesian semi-supervised learning with graph gaussian processes. *Advances in Neural Information Processing Systems*, 31, 2018.
- [54] Vu-Linh Nguyen, Sébastien Destercke, and Eyke Hüllermeier. Epistemic uncertainty sampling. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pages 72–86. Springer, 2019.
- [55] CAZHAOW SALIH OAZA. Bayesian error bars for regression. 1996.
- [56] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, et al. The pagerank citation ranking: Bringing order to the web. 1999.
- [57] John Palowitch, Anton Tsitsulin, Brandon Mayer, and Bryan Perozzi. Graphworld: Fake graphs bring real insights for gnns. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3691–3701, 2022.
- [58] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.
- [59] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

- [60] Adrian Schwaiger, Poulami Sinhamahapatra, Jens Gansloser, and Karsten Roscher. Is uncertainty quantification in deep learning sufficient for out-of-distribution detection? *Aisafety@ijcai*, 54, 2020.
- [61] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29:93–93, 2008.
- [62] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.
- [63] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.
- [64] Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021.
- [65] Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.
- [66] Jayaraman Thiagarajan, Rushil Anirudh, Vivek Sivaraman Narayanaswamy, and Timo Bremer. Single model uncertainty estimation via stochastic data centering. *Advances in Neural Information Processing Systems*, 35:8662–8674, 2022.
- [67] Puja Trivedi, Mark Heimann, Rushil Anirudh, Danai Koutra, and Jayaraman J Thiagarajan. Accurate and scalable estimation of epistemic uncertainty for graph neural networks. *arXiv preprint arXiv:2401.03350*, 2024.
- [68] Dennis Thomas Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. 2021.
- [69] Joost Van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.
- [70] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.
- [71] Fangxin Wang, Yuqing Liu, Kay Liu, Yibo Wang, Sourav Medya, and Philip S Yu. Uncertainty in graph neural networks: A survey. *arXiv preprint arXiv:2403.07185*, 2024.
- [72] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.
- [73] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation. 2022.
- [74] Min Wang, Hao Yang, Jincai Huang, and Qing Cheng. Moderate message passing improves calibration: A universal way to mitigate confidence bias in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21681–21689, 2024.
- [75] Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration, 2022.
- [76] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11*, page 681–688, Madison, WI, USA, 2011. Omnipress.
- [77] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

- [78] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [79] Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*, 2023.
- [80] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.
- [81] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1100–1109, New York, New York, USA, 20–22 Jun 2016. PMLR.
- [82] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33:12827–12836, 2020.
- [83] Yin-Cong Zhi, Yin Cheng Ng, and Xiaowen Dong. Gaussian processes on graphs via spectral kernel learning. *IEEE Transactions on Signal and Information Processing over Networks*, 2023.
- [84] Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. Dirichlet energy constrained learning for deep graph neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21834–21846. Curran Associates, Inc., 2021.

A Proofs

A.1 EBM for i.i.d. Data

We first consider proofs for formal statements regarding logit-based EBMs in general (i.e. for i.i.d. data).

Proposition 4.1. *Let $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^C$ be a piecewise affine function and $\mathbb{R}^h = \bigcup_l^L Q_l$ be the disjoint set of polytopes on which f_θ is affine, i.e. $f_\theta(\mathbf{x}) = \mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)}$ for $\mathbf{x} \in Q_l$. Assuming the direction of the rows of each $\mathbf{W}^{(l)}$ to be uniformly distributed, the probability that Z_θ converges decreases exponentially in the number of non-closed linear regions L' and classes C .*

$$\Pr[Z_\theta < \infty] \approx (1/2)^{C \cdot L'}$$

Proof. Consider an arbitrary direction $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = 1$. As shown in [28] (Lemma 3.1), there exists some t and α_0 such that for all $\alpha > \alpha_0$ it holds that $\alpha\mathbf{x} \in Q_t$. We have $f_\theta(\alpha\mathbf{x}) = \mathbf{W}^{(t)}\alpha\mathbf{x} + \mathbf{b}^{(t)}$, and therefore:

$$-E_\theta(\alpha\mathbf{x}, y) = \mathbf{W}_y^{(t)}\alpha\mathbf{x} + \mathbf{b}_y^{(t)}$$

We now distinguish between two cases as $\alpha \rightarrow \infty$:

If $\mathbf{W}_y^{(t)}\mathbf{x} \geq 0$, we have:

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \alpha \mathbf{W}_y^{(t)}\mathbf{x} &= \infty \\ \lim_{\alpha \rightarrow \infty} \alpha \mathbf{W}_y^{(t)}\mathbf{x} + \mathbf{b}_y^{(t)} &= \infty \\ \lim_{\alpha \rightarrow \infty} E_\theta(\alpha\mathbf{x}, y) &= -\infty \end{aligned}$$

And conversely if $\mathbf{W}_y^{(t)}\mathbf{x} \leq 0$, we have:

$$\lim_{\alpha \rightarrow \infty} E_\theta(\alpha\mathbf{x}, y) = \infty$$

Since we assumed the direction of $\mathbf{W}_y^{(t)}$ to be uniformly distributed, both cases occur with probability 0.5. Now we consider the marginal energy assigned to $\alpha\mathbf{x}$:

$$\begin{aligned} E_\theta(\alpha\mathbf{x}) &= -\log \sum_y \exp(-E_\theta(\alpha\mathbf{x}, y)) \\ &\leq \alpha \min_y \{-E_\theta(\mathbf{x}, y)\} \end{aligned}$$

We now analyze when the marginal energy $E_\theta(\alpha\mathbf{x})$ diverges toward ∞ and $-\infty$ respectively. If for any $y \in \{1, \dots, C\}$ we have that $\mathbf{W}_y^{(t)}\mathbf{x} \leq 0$, then:

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} E_\theta(\alpha\mathbf{x}) &\leq \lim_{\alpha \rightarrow \infty} \alpha \min_y \{E_\theta(\mathbf{x}, y)\} \\ &= -\infty \end{aligned}$$

Since each $\mathbf{W}_y^{(t)}\mathbf{x} > 0$ with probability 1/2, the limit diverges toward $-\infty$, with probability at most $(1/2)^C$.

Now consider some open set $Q^{(l)}$ with a non-zero measure for over which the classifier f_θ is linear. With probability at least 1/2, for $\mathbf{x} \in Q^{(l)}$, f_θ diverges toward ∞ .

We now can consider the integral over the unnormalized density implied by f_θ :

$$\begin{aligned} Z_\theta &= \int_{\mathbb{R}^d} \exp(-E_\theta(\mathbf{x})) d\mathbf{x} \\ &= \sum_l \int_{Q^{(l)}} \exp(-E_\theta(\mathbf{x})) d\mathbf{x} \end{aligned}$$

For any open region $Q^{(l)}$, the energy will diverge toward $-\infty$ with probability $(1/2)^C$ and therefore:

$$\int_{Q^{(l)}} \exp(-E_\theta(\mathbf{x})) d\mathbf{x} = \infty$$

For Z_θ to be finite and well-defined, the integral over all L' open linear regions with non-zero measures need to converge simultaneously, each of which happens independently with probability $(1/2)^C$. Therefore:

$$\Pr[Z_\theta < \infty] = (1/2)^{C \cdot L'}$$

□

Remark. Note that the assumption that the rows of each $\mathbf{W}^{(l)}$ have a uniformly distributed direction is given when initializing the model from a normal distribution. When training a classifier with the commonly used cross-entropy objective, the training incentivizes the model to assign high logits to exactly one of the classes for each datapoint. Therefore, the model is intuitively encouraged to have at least one row of $\mathbf{W}^{(l)}$ for which the $\mathbf{x}$ that are in the corresponding region $Q^{(l)}$ have positive projection $\mathbf{W}_y^{(l)} \mathbf{x} > 0$. In practice, we expect that for a trained model the probability of any $\mathbf{x}$ to have a negative projection onto all rows of the corresponding $\mathbf{W}^{(l)}$ to be lower than $(1/2)^C$.

Theorem 4.2. For a piecewise affine classifier f_θ as in Proposition 4.1, $\hat{p}_\theta$ is well-defined.

$$\hat{Z}_\theta = \int_{\mathcal{X}} \sum_y \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(\mathbf{f}_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)^\gamma d\mathbf{x} < \infty$$

Proof. We consider the corrected joint energy $\hat{E}_\theta(\alpha \mathbf{x}, y)$ for some sufficiently large $\alpha > \alpha_0$ and an arbitrary direction $\mathbf{x} \in \mathbb{R}^d$.

$$\begin{aligned} \hat{E}_\theta(\alpha \mathbf{x}, y) &= E_\theta(\alpha \mathbf{x}, y) - \gamma \log \mathcal{N}(\mathbf{f}_\theta(\alpha \mathbf{x})_y \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \\ &= -f_\theta(\alpha \mathbf{x})_y + \frac{\gamma}{2} (\mathbf{f}_\theta(\alpha \mathbf{x})_y - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1} (\mathbf{f}_\theta(\alpha \mathbf{x})_y - \boldsymbol{\mu}_y) + \text{const} \\ &= -\mathbf{W}_y^{(l)} \alpha \mathbf{x} - b_y^{(l)} + \frac{\gamma}{2} (\mathbf{W}_y^{(l)} \alpha \mathbf{x} + b_y^{(l)} - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1} (\mathbf{W}_y^{(l)} \alpha \mathbf{x} + b_y^{(l)} - \boldsymbol{\mu}_y) + \text{const} \end{aligned}$$

For large enough α_0 , the terms quadratic in $\alpha \mathbf{x}$ will dominate linear terms and we can write for some $c > 0$:

$$\begin{aligned} \hat{E}_\theta(\alpha \mathbf{x}, y) &\geq c * (\mathbf{W}_y^{(l)} \alpha \mathbf{x})^T \boldsymbol{\Sigma}_y^{-1} (\mathbf{W}_y^{(l)} \alpha \mathbf{x}) \\ &= \alpha^2 (\sqrt{c} \mathbf{W}_y^{(l)} \mathbf{x})^T \boldsymbol{\Sigma}_y^{-1} (\sqrt{c} \mathbf{W}_y^{(l)} \mathbf{x}) \\ &\geq \mathbf{x}^T \boldsymbol{\Sigma}_*^{-1} \mathbf{x} \end{aligned}$$

Here, we pick $\boldsymbol{\Sigma}_*$ such that the equation holds for all $y \in \{1, \dots, C\}$. Note that this means that in the limit, each joint energy term is bounded by the logarithm of the same Gaussian $\log \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_*)$. For the marginal energy it holds:

$$\begin{aligned}
\hat{E}_\theta(\alpha \mathbf{x}) &= -\log \sum_y \exp \left(-\hat{E}_\theta(\alpha \mathbf{x}, y) \right) \\
&\geq -\log \sum_y \exp \left(-\alpha \mathbf{x}^T \boldsymbol{\Sigma}_*^{-1} \alpha \mathbf{x} \right) \\
&= \alpha \mathbf{x}^T \boldsymbol{\Sigma}_*^{-1} \alpha \mathbf{x} - \log C \\
&\geq \alpha \mathbf{x}^T \boldsymbol{\Sigma}'_*{}^{-1} \alpha \mathbf{x}
\end{aligned}$$

Here, we absorb the constant $\log C$ into the quadratic term. Similar to the proof of Proposition 4.1, we now analyze the integral of the implied density for each linear region of the classifier $Q^{(l)}$.

$$\int_{Q^{(l)}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} = \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 \leq \alpha_0}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} + \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 > \alpha_0}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x}$$

The first term is an integral of a finite function over a finite region, which therefore will also be finite:

$$I_{\alpha \leq \alpha_0}^{(l)} := \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 \leq \alpha_0}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} < \infty$$

For the second term, we notice that it is bounded by the (finite) Gaussian integral:

$$\begin{aligned}
I_{\alpha > \alpha_0}^{(l)} &:= \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 > \alpha_0}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} \\
&\leq \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 > \alpha_0}} \exp \left(-\mathbf{x}^T \boldsymbol{\Sigma}'_*{}^{-1} \mathbf{x} \right) d\mathbf{x} \\
&< \infty
\end{aligned}$$

Lastly, since we can decompose $\mathbf{R}^d$ into a finite set of linear regions over each of which the integral is finite, the entire integral is finite.

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} &= \sum_{l=1}^L \int_{Q^{(l)}} \exp \left(-\hat{E}_\theta(\mathbf{x}) \right) d\mathbf{x} \\
&= \sum_{l=1}^L \left(I_{\alpha \leq \alpha_0}^{(l)} + I_{\alpha > \alpha_0}^{(l)} \right) \\
&< \infty
\end{aligned}$$

□

Theorem A.1.1. For a piecewise affine classifier f_θ as in Proposition 4.1, and a piecewise affine feature extractor $g_\theta(\mathbf{x})$, $\hat{p}_\theta$ is integrable.

$$\hat{Z}_\theta = \int_{\mathcal{X}} \sum_y \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(g_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)^\gamma d\mathbf{x} < \infty$$

Proof. Note that both affine functions f_θ and g_θ partition the input space $\mathbb{R}^d$ into a finite set of linear regions. Intersecting their boundaries again gives rise to a finite set of linear regions. For bounded

regions, the integral over the induced density will be finite, so we discuss the unbounded regions here. For any such region $Q^{(l)}$, we have $f_\theta(\mathbf{x}) = \mathbf{w}_f^T \mathbf{x} + \mathbf{b}_f$ and $g_\theta(\mathbf{x}) = \mathbf{W}_g \mathbf{x} + \mathbf{b}_g$.

We now partition the $\mathbb{R}^d \cap Q^{(l)}$ into four regions using the kernel spaces of $\mathbf{w}_f$ and $\mathbf{W}_g$ respectively, i.e. $\mathbb{R}^d \cap Q^{(l)} = \mathbf{0}_{f+g} \cup \mathbf{0}_{f-g} \cup \mathbf{0}_{g-f} \cup \mathbf{I}_{f+g}$, where we define $\mathbf{0}_{f+g} := \ker(\mathbf{w}_f) \cap \ker(\mathbf{W}_g) \cap Q^{(l)}$, $\mathbf{0}_{f-g} := \ker(\mathbf{w}_f) \setminus \ker(\mathbf{W}_g) \cap Q^{(l)}$, $\mathbf{0}_{g-f} := \ker(\mathbf{W}_g) \setminus \ker(\mathbf{w}_f) \cap Q^{(l)}$ and $\mathbf{I}_{f+g} := \text{im}(\mathbf{w}_f) \cap \text{im}(\mathbf{W}_g) \cap Q^{(l)}$.

We can now decompose the integral over $Q^{(l)}$ into integrals over all four regions. Note that of these sets, only $\mathbf{I}_{f+g}$ has a non-zero measure (if the affine classifier is not the null function). Therefore, we only have to focus on this domain. For some α_0 and $\mathbf{x} \in \mathbf{I}_{f+g}$, $\|\mathbf{x}\|_2 \geq \alpha_0$, we have:

$$\begin{aligned} \hat{p}_\theta(\mathbf{x}) &\propto \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(g_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \\ &\propto \exp(\mathbf{w}_f^T \mathbf{x} + \mathbf{b}_f) \exp(-(\mathbf{W}_g \mathbf{x} + \mathbf{b}_g)^T \boldsymbol{\Sigma}_g^{-1} (\mathbf{W}_g \mathbf{x} + \mathbf{b}_g)) \\ &\leq \exp(-\mathbf{x}^T \boldsymbol{\Sigma}_{*,y}^{-1} \mathbf{x}) \end{aligned}$$

For some $\boldsymbol{\Sigma}_{*,y}^{-1}$ that bounds the quadratic function in the exponential for $\|\mathbf{x}\|_2 \geq \alpha_0$. The integral over the region $Q^{(l)}$ now reduces to:

$$\begin{aligned} I^{(l)} &= \int_{\mathbf{x} \in Q^{(l)}} \sum_y \text{const} * \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(g_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) d\mathbf{x} \\ &= \text{const} + \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 \geq \alpha_0}} \sum_y \text{const} * \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(g_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) d\mathbf{x} \\ &\leq \text{const} + \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 \geq \alpha_0}} \sum_y \text{const} * \exp(-\mathbf{x}^T \boldsymbol{\Sigma}_{*,y}^{-1} \mathbf{x}) d\mathbf{x} \\ &\leq \text{const} + \int_{\substack{Q^{(l)} \\ \|\mathbf{x}\|_2 \geq \alpha_0}} \text{const} * \exp(-\mathbf{x}^T \boldsymbol{\Sigma}_*^{-1} \mathbf{x}) d\mathbf{x} \\ &< \infty \end{aligned}$$

Here, we pick $\boldsymbol{\Sigma}_*$ to bound the quadratic terms of all $\boldsymbol{\Sigma}_{y,*}$. As previously mentioned, the kernel spaces of $\mathbf{w}_f$ and $\mathbf{W}_g$ have zero measure and therefore do not contribute to the integral. As previously, the integral over the entire domain $\mathbb{R}^d$ decomposes into finite integrals over all $Q^{(l)}$.

$$\begin{aligned} \int_{\mathbb{R}} \sum_y \text{const} * \exp(-E_\theta(\mathbf{x}, y)) \mathcal{N}(g_\theta(\mathbf{x}) \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) d\mathbf{x} &= \text{const} + \sum_{l=1}^L I^{(l)} \\ &< \infty \end{aligned}$$

□

Corollary 4.3. For a piecewise affine classifier f_θ as in Proposition 4.1, and any $\mathbf{x} \in \mathbb{R}^d$ almost surely:

$$\lim_{\alpha \rightarrow \infty} \hat{p}_\theta(\alpha \mathbf{x}) = 0$$

Proof. As per the proof of Theorem 4.2, for sufficiently large α we have for some $\boldsymbol{\Sigma}_*^{-1}$:

$$\hat{E}_\theta(\alpha \mathbf{x}) \geq \mathbf{x}^T \boldsymbol{\Sigma}_*^{-1} \mathbf{x}$$

From there, it follows directly that:

$$\begin{aligned}
 \lim_{\alpha \rightarrow \infty} \hat{p}_\theta(\alpha \mathbf{x}) &= \lim_{\alpha \rightarrow \infty} \exp(-\hat{E}_\theta(\alpha \mathbf{x})) \\
 &= \exp\left(\lim_{\alpha \rightarrow \infty} -\hat{E}_\theta(\alpha \mathbf{x})\right) \\
 &\leq \exp\left(\lim_{\alpha \rightarrow \infty} -\mathbf{x}^T \Sigma_*^{-1} \mathbf{x}\right) \\
 &= 0
 \end{aligned}$$

□

A.2 Graph-based Energies at Different Scales

Here, we show that local energy (Equation (5)) and group energy (Equation (6)) induce a valid probability density when using a linear diffusion operator P_A and regularized energy according to Equation (8). Note that independent energy (Equation (4)) is shown to induce a well-defined probability density as it is just a graph-independent regularized energy.

Proposition A.2.1. *For a linear diffusion operator $P_A(\mathbf{x}) = \alpha \mathbf{x} + \text{const}$, $\alpha > 0$ and the regularized energy $\hat{E}_\theta(\mathbf{x}, y)$, the local energy $\hat{E}_L(\mathbf{x})$ induces a well-defined density:*

$$\int_{\mathcal{X}} \exp(-\hat{E}_{\theta,L}(\mathbf{x})) d\mathbf{x} < \infty$$

Proof.

$$\begin{aligned}
 \hat{E}_{\theta,L}(\mathbf{x}) &= -\log \sum_y \exp\left(P_A\left(-\hat{E}_\theta(\mathbf{x}, y)\right)\right) \\
 &= -\log \sum_y \exp\left(-\alpha \hat{E}_\theta(\mathbf{x}, y) + \text{const}\right) \\
 &\geq -\log \sum_y \exp\left(-\alpha \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \text{const}\right) \\
 &= \alpha \mathbf{x}^T \Sigma^{-1} \mathbf{x} + \text{const} + \log C \\
 &\geq \mathbf{x}^T \Sigma'^{-1} \mathbf{x}
 \end{aligned}$$

Again, we absorbed the constant terms as well as α into the quadratic term for large enough $\|\mathbf{x}\|_2$ and used the quadratic bound from Theorem 4.2. From here, it is straightforward that:

$$\int_{\mathcal{X}} \exp\left(-\hat{E}_{\theta,L}(\mathbf{x})\right) d\mathbf{x} \leq \text{const} + \int_{\mathcal{X}} \exp\left(-\mathbf{x}^T \Sigma'^{-1} \mathbf{x}\right) d\mathbf{x} < \infty$$

□

Proposition A.2.2. *For a linear diffusion operator $P_A(\mathbf{x}) = \alpha \mathbf{x} + \text{const}$, $\alpha > 0$ and the regularized energy $\hat{E}_\theta(\mathbf{x}, y)$, the group energy $\hat{E}_G(\mathbf{x})$ induces a well-defined density:*

$$\int_{\mathcal{X}} \exp(-\hat{E}_{\theta,G}(\mathbf{x})) d\mathbf{x} < \infty$$

Proof.

$$\begin{aligned}
 \hat{E}_{\theta,G}(\mathbf{x}) &= P_A\left(\hat{E}_\theta(\mathbf{x})\right) \\
 &= \alpha \hat{E}_\theta(\mathbf{x}) + \text{const} \\
 &\geq \alpha \mathbf{x}^T \Sigma_*^{-1} \mathbf{x} + \text{const} \\
 &\geq \mathbf{x}^T \Sigma'^{-1} \mathbf{x}
 \end{aligned}$$

Again, we have used the quadratic bound of Theorem 4.2 and absorbed $\alpha > 0$ for large enough $\|\mathbf{x}\|_2$. From there, it follows:

$$\int_{\mathcal{X}} \exp\left(-\hat{E}_{\theta,G}(\mathbf{x})\right) d\mathbf{x} \leq \text{const} + \int_{\mathcal{X}} \exp\left(-\mathbf{x}^T \Sigma'^{-1}_* \mathbf{x}\right) d\mathbf{x} < \infty$$

□

Remark. We want to point out that most diffusion operations, especially the ones discussed in Appendix C.6 including the label-propagation smoothing used in our experiments, are of the form $P_A(\mathbf{x}) = \alpha\mathbf{x} + \text{const}$ with $\alpha > 0$. Since we only integrate over the features of a single node and keep all other features fixed, all of these diffusion processes can be expressed as a weighted sum over nodes in the graph. Since we consider unweighted graphs, there are no negative edge weights which ensures that $\alpha > 0$ (as long as some sort of self-loop is included). Therefore, linear diffusion processes can be seen as positive affine transformations.

Theorem 4.4. For a linear diffusion operator $P_A(\mathbf{x}) = \alpha\mathbf{x} + \text{const}$, $\alpha > 0$ and the regularized energy $\hat{E}_{\theta}(\mathbf{x}, y)$, GEBM induces a well-defined density:

$$\int_{\mathcal{X}} \exp\left(-\hat{E}_{\theta,\text{GEBM}}(\mathbf{x})\right) d\mathbf{x} < \infty$$

Proof. We previously established quadratic bounds on all three constituents of the aggregate energy for large enough $\|\mathbf{x}\|_2$:

$$\begin{aligned}\hat{E}_{\theta,I}(\mathbf{x}) &\geq \mathbf{x}^T \Sigma_I^{-1} \mathbf{x} \\ \hat{E}_{\theta,L}(\mathbf{x}) &\geq \mathbf{x}^T \Sigma_L^{-1} \mathbf{x} \\ \hat{E}_{\theta,G}(\mathbf{x}) &\geq \mathbf{x}^T \Sigma_G^{-1} \mathbf{x}\end{aligned}$$

Hence, there must be one $\Sigma_*^{-1} \in \{\Sigma_I^{-1}, \Sigma_L^{-1}, \Sigma_G^{-1}\}$ that bounds all three energy types. We now look at the GEBM model:

$$\begin{aligned}\hat{E}_{\theta,\text{GEBM}}(\mathbf{x}) &= \log\left(\exp\left(\hat{E}_{\theta,I}(\mathbf{x})\right) + \exp\left(\hat{E}_{\theta,L}(\mathbf{x})\right) + \exp\left(\hat{E}_{\theta,G}(\mathbf{x})\right)\right) \\ &\geq \log\left(3 * \exp\left(\mathbf{x}^T \Sigma_*^{-1} \mathbf{x}\right)\right) \\ &= \mathbf{x}^T \Sigma_*^{-1} \mathbf{x} + \log 3 \\ &\geq \mathbf{x}^T \Sigma'^{-1}_* \mathbf{x}\end{aligned}$$

As before, this bounds the density induced by GEBM:

$$\int_{\mathcal{X}} \exp\left(-\hat{E}_{\theta,\text{GEBM}}(\mathbf{x})\right) d\mathbf{x} \leq \text{const} + \int_{\mathcal{X}} \exp\left(-\mathbf{x}^T \Sigma'^{-1}_* \mathbf{x}\right) d\mathbf{x} < \infty$$

□

Remark. This directly imposes a restriction on which energies can be included in GEBM beyond the three naturally arising graph-specific terms we propose: As long as the energy term can be bounded by an energy that grows fast enough to ensure convergence, our framework accommodates it.

We also can explicitly write out the (unnormalized) GEBM density $p_{\theta,\text{GEBM}}(\mathbf{x})$ in terms of its constituents $p_{\theta,I}(\mathbf{x})$, $p_{\theta,L}(\mathbf{x})$ and $p_{\theta,G}(\mathbf{x})$.

$$p_{\theta,\text{GEBM}}(\mathbf{x}) \propto \frac{1}{p_{\theta,I}(\mathbf{x})^{-1} + p_{\theta,L}(\mathbf{x})^{-1} + p_{\theta,G}(\mathbf{x})^{-1}}$$

B Experimental Setup

B.1 Datasets and Distribution Shifts

Dataset	#Nodes n	#Edges m	#Features d	#Classes c	Avg. Feature Density (%)	Homophily (%)	Edge Density m/n^2 (%)	Left-out-Classes	#Nodes-i.d. (LoC) n_{id}
CoraML	2995	16316	2879	7	1.75	78.9	0.18	3	1650
CoraML LLM	2995	16316	384	7	n.a.	78.9	0.18	3	1650
Citeseer	4230	10674	602	6	0.76	94.9	0.06	2	2977
PubMed	19717	88648	500	3	n.a.	80.2	0.02	1	11842
Amazon Photo	7650	238162	745	8	34.74	82.7	0.41	3	4555
Amazon Computers	13752	491722	767	10	34.84	77.7	0.26	4	10000
Coauthor CS	18333	163788	6805	15	0.88	80.8	0.05	5	9424
Coauthor Physics	34493	495924	8415	5	0.39	93.1	0.04	2	28221

Table 5: Statistics about the datasets used in this work.

We use eight datasets in this work that we expose to similar kinds of distribution shifts. For CoraML, we also create a version that uses continuous word embeddings instead of categorical bag-of-word features. We use the *all-MiniLM-L6-v2* sentence transformer provided by Hugging Face [78] to embed the abstracts that are provided for each paper (node) in the citation network. All datasets are taken from PyTorch Geometric [19]. They are licensed as C.C.0 1.0 (CoraML, CoraML LLM), C.C. Attribution-NonCommercial-Share Alike 3.0 (Citeseer), OdbL 1.0 (PubMed).

We expose each dataset to the same distribution shifts which can be categorized into three families:

- (i) **Leave-out-Classes.** We pre-select a subset of classes and designate them as out-of-distribution. That is, we remove them from the training set and train the GNN on the remaining set of nodes. We focus on an inductive setting, where we also remove all edges linking o.o.d. left-out-class nodes to the training graph, thus preventing information leakage. In the transductive setting, o.o.d. nodes still contribute to the training signal as they may be connected to i.d. nodes. This distribution shift can be seen as occurring on a cluster level, as we introduce anomalous nodes that are likely to cluster together. We distinguish between two kinds of selection processes for the classes to be left out: We either pick the last classes in the dataset (*LoC (last)*) or choose classes with the most heterophilic connection pattern, which distinguishes them even further from i.d. classes (*LoC (hetero.)*).
- (ii) **Feature Perturbations.** We select 50% of nodes at random to be o.o.d. and perturb their features by replacing them with random noise. We distinguish between three perturbation types that control the similarity of node features to training data: We generate features that are similar to i.d. data (*near-o.o.d.*) in the following way: For each of the d features, we compute how frequent it occurs in the dataset. We then proceed to sample each of these features independently with the corresponding success probability $\hat{p}$ from a Bernoulli $\text{Ber}(\hat{p})$. Instead, we can also set the success probability to $p = 0.5$ for each of the features and induce a more severe distribution shift within the bag-of-words domain of the dataset. Lastly, when sampling features according to a normal distribution $\mathcal{N}(0, 1)$, we generate out-of-domain data that should, in theory, be easy to detect *far-o.o.d.* For the CoraML-LM and the PubMed datasets, features are not bag-of-words. Therefore, the near-o.o.d. shift is omitted and the far-o.o.d. shift needs to be considered within the domain of the data.
- (iii) **Structural.** We induce structure-related shifts in two ways: The first option we consider is to rank all nodes according to their local homophily. That is, we compute the ratio of neighbors with the same class: $h_i = |\{v_j \in \mathcal{N}_i : y_i = y_j\}|/|\mathcal{N}_i|$. We then designate 50% of nodes with the highest heterophily (i.e. lowest homophily) as o.o.d. (*homophily*). The second structural shift ranks nodes according to their (approximate) Page Rank centrality and declares nodes with low values as o.o.d. (*Page Rank*).

In the inductive setting, we remove the o.o.d. nodes from the training graph and re-introduce them during inference. We evaluate o.o.d. detection metrics on a validation/test set that includes both i.d. and o.o.d. nodes. While the training and validation set are randomized for each split, the test set is shared across all splits to prevent data leakage. All results are reported over five different splits and five independent model weight initializations for each of them. Where appropriate, we also report standard deviations in Appendix C.

Remarks regarding Page Rank Shifts. We study the centrality-based shift (which heavily correlates with a degree-based shift) for completeness reasons as a similar generation process is used in [79]. However, to which extent methods should be expected to assign high uncertainty to low-degree nodes is questionable: First, centrality and/or degree are feature-irrespective quantities. In contrast to homophily (nodes of similar classes have similar features), the confidence of a classifier may not (and should not) depend on the node degree/centrality. Furthermore, many baselines use symmetric normalization (Appendix C.6) that inherently favors high-degree nodes. Therefore, an evaluation regarding node degree/centrality as a distribution shift will not test the quality of an uncertainty measure but instead favor all models that use appropriate diffusion processes for most baselines. Regardless of our concerns regarding a centrality-based shift, we report results regarding that shift as well in Appendix C. Note that while GEBM is not among the best-performing estimators in this setting, it still achieves the highest average o.o.d. detection rank overall, indicating that its merits in all other settings heavily outweigh its performance on the Page Rank shift.

B.2 Models and Training

At the backbone of all models, we use the same GCN [34] architecture if not specified explicitly otherwise. We use one hidden layer of dimension 64, symmetric normalization Appendix C.6, and add self-loops to the undirected (symmetric) adjacency matrix. We use ReLU nonlinearities, enable the bias term, and use dropout at $p = 0.5$.

All models are trained with the ADAM optimizer [33] with a learning rate of 10^{-3} , weight decay of 10^{-4} , and a cross-entropy objective. We use early stopping on the validation loss with a patience of 50, an absolute improvement threshold of 10^{-1} , and select the model with the best validation loss. We implement our models in PyTorch [58] and PyTorch Geometric [19] and train on two types of machines: (i) Xeon E5-2630 v4 CPU @ 2.20GHz with a NVIDIA GTX 1080TI GPU and 128 GB of RAM. (ii) AMD EPYC 7543 CPU @ 2.80GHz with a NVIDIA A100 GPU and 128 GB of RAM.

As the GCN backbone used in our experiments is lightweight, model training finishes within a few minutes and VRAM consumption is dominated by the datasets. Our post hoc method can be fitted and evaluated in negligible time ($<1s$).

For MC-Dropout [22], we use a dropout probability of $p = 0.5$ which we also use for DropEdge [59]. At inference, we evaluate 50 samples to compute uncertainty. For ensembles, we train 10 backbones from different weight initializations independently. The Bayesian GNN [6, 15, 18] also is evaluated with 50 samples during inference and uses a KL-loss with weight 10^{-1} and is trained with learning rate 10^{-2} and no weight-decay. We parametrize the weight distribution using a log transform and initialize the mean and log scale to 1.0 and -3.0 respectively. Weight distributions are regularized to follow a standard normal.

For GPN [64], we follow the hyperparameters suggested by the authors and use warmup training on the normalizing flow for 5 epochs with a learning rate of $1e-2$, no weight decay during joint training and $1e-2$ during warmup. The flow dimension is 16 and is composed of 10 radial layers. The cross-entropy regularization weight is 10^{-4} . We use Page Rank propagation for 10 iterations at a teleport probability of 0.1.

For SGCN [82], we use the same GCN backbone and a GDK-prior at cutoff distance 10 and scale $\sigma = 1.0$. For teacher training, we use a learning rate of 10^{-2} , weight decay of $5 * 10^{-4}$ and a KL-loss weight of 10^{-1} .

For HEAT [36], we use the hyperparameters suggested by the authors. In contrast to their study on the image domain, we do not have features with spatial extent and can not use standard deviation pooling on the volume. We set the temperature parameter of the combined HEAT model to -1 , following the suggestion of the authors. Similar to their EBM backbone, we imitate the structure of the classifier and use a 2-layer MLP with a hidden dimension of 64 akin to the GCN backbone.

Our GEBM framework regularizes logit-based joint energy at a strength γ . For downstream tasks, this parameter can, in general, be tuned. We find that an equal weighting of predictive energy and regularization performs well: We choose γ such that the 95% quantiles of both the training logits and the representations on which the Gaussian density model is fit are in the same range. Furthermore, recent work on deterministic uncertainty argues that density-based epistemic uncertainty should not be estimated directly from the logits [46]. We follow this advice and do not fit and evaluate the

regularizer on the output layer (logits) of the GNN but instead on its penultimate layer representations. Note that for the ReLU networks we use, all formal proofs still hold in this setting as the penultimate representation of a node v is also described by a piecewise affine function [28] (see A.1.1). As a smoothing operator P_A , we employ label-propagation smoothing with $\alpha = 0.5$ for $t = 10$ iterations. Lastly, we found that aggregating the energy terms at different scales (Equation (9)) using a sum operation instead of logsumexp to perform better in practice and use it for our experiments.

In our ablation regarding different model backbones, we use the default hyperparameters from the GCN backbone (e.g. number of hidden layers). For each model additional hyperparameters are set to the following values: For GATv2 [70, 8] we use 8 heads and use summation to aggregate them. For SAGE [27], we use no normalization at the layers.

B.3 Uncertainty Evaluation

For sampling-based approaches, we compute aleatoric and epistemic uncertainty as entropy and mutual information respectively according to [24]:

$$u^{\text{alea}} = \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})} [\mathbb{H}[\mathbf{p}(y | \mathbf{x})]]$$

$$u^{\text{epi}} = \mathbb{MII}[\theta, \mathbf{y} | \mathbf{x}, \mathcal{D}] = \mathbb{H}[\mathbb{E}_{\theta \sim p(\theta|\mathcal{D})} [\mathbf{p}(y | \mathbf{x})]] - \mathbb{E}_{\theta \sim p(\theta|\mathcal{D})} [\mathbb{H}[\mathbf{p}(y | \mathbf{x})]]$$

For evidential approaches, we follow [64] and use the maximum softmax response $\max_c \mathbf{p}(y = c | \mathbf{x})$ as aleatoric uncertainty and the total evidence $\sum_c \alpha_c$ as an epistemic estimate. Since we want to compare single measures of epistemic uncertainty, we evaluate GPN's epistemic uncertainty *in the presence of network effects*. For deterministic models (that are at the backbone of EBM-based approaches), we use the predictive entropy as a measure of aleatoric uncertainty $\mathbb{H}[\mathbf{p}(y | \mathbf{x})]$ and the energy $E_\theta(\mathbf{x})$ as a measure of epistemic uncertainty. For all EBMs, we use a temperature of $\tau = 1.0$.

C Additional Results

C.1 Out-of-Distribution Detection

We report AUC-ROC and AUC-PR metrics for the o.o.d. detection problem in an inductive and transductive setting with corresponding standard deviations over all five splits and five model initializations each in Tables 6, 8, 10 and 12. Again, we report the average performance ranks are in Tables 2, 9, 11 and 13. As mentioned in Section 5.2, we consider a weighted average that does not favor any distribution shift family. That is, we assign a weight of 1/6 to structural and leave-out-class settings and factor in feature perturbations at a weight of 1/9 each.

In an inductive setting, GEBM outperforms all other baselines and achieves the best rank regarding both AUC-ROC and AUC-PR metrics. In the transductive setting, the two evidential methods GPN and SGCN outperform GEBM on two datasets: We argue that this is due to feature leakage as o.o.d. data is present during training and evidential models are explicitly encouraged to assign low evidence to anomalous nodes. We want to point out that it is unrealistic to assume that o.o.d. data is available in practice and therefore strongly argue in favor of the inductive scenario as a more realistic benchmark. Nonetheless, GEBM is highly effective for transductive problems as well.

C.2 Improvement over Second Best Method

We evaluate the improvement in AUC-ROC scores of GEBM over the estimator assigned the overall second-best rank. To that end, we average the model ranks not individually for each estimator and dataset, but instead compute an average over datasets and splits simultaneously. This way, we obtain a global rank (over datasets and shifts) for each epistemic estimate. Again, to not favor certain classes of shifts, we assign weights such that each distribution shift family has an equal contribution. Based on AUC-ROC scores listed in Table 6, we list the rank of each model in Table 14:

While our approach, GEBM, also ranks the highest globally, a vanilla logit-based EBM is the next best approach. Therefore, we compute the improvement in AUC-ROC scores over this model achieved by

	Model	LoC (<i>latent</i>)		LoC (<i>hetero</i>)		Ber(<i>p</i>) (<i>near</i>)		Ber(0.5)		N(0,1) (<i>far</i>)		Page Rank		Homophily	
		AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$	AUC-ROC $\uparrow$	Acc. $\uparrow$
CoNLL	GCN-MCD	89.7 $\pm$ 1.4/99.9 $\pm$ 0.2	90.4 $\pm$ 0.2	87.8 $\pm$ 1.5/90.3 $\pm$ 0.2	87.4 $\pm$ 0.2	66.4 $\pm$ 1.4/49.2 $\pm$ 0.2	75.6 $\pm$ 0.2	58.9 $\pm$ 1.5/50.4 $\pm$ 0.2	75.6 $\pm$ 0.2	31.4 $\pm$ 1.4/48.7 $\pm$ 0.2	75.6 $\pm$ 0.2	72.9 $\pm$ 1.5/50.2 $\pm$ 0.2	82.3 $\pm$ 0.2	71.4 $\pm$ 1.5/50.2 $\pm$ 0.2	91.4 $\pm$ 0.2
	GCN-DE	86.2 $\pm$ 1.7/73.0 $\pm$ 0.1	87.2 $\pm$ 0.1	83.1 $\pm$ 1.5/66.8 $\pm$ 0.2	82.0 $\pm$ 0.2	47.1 $\pm$ 1.4/68.8 $\pm$ 0.2	71.9 $\pm$ 0.2	47.1 $\pm$ 1.4/68.8 $\pm$ 0.2	71.9 $\pm$ 0.2	26.8 $\pm$ 1.5/57.8 $\pm$ 0.2	71.9 $\pm$ 0.2	66.9 $\pm$ 1.5/69.7 $\pm$ 0.2	79.1 $\pm$ 0.2	70.8 $\pm$ 1.5/64.5 $\pm$ 0.2	85.5 $\pm$ 0.2
	GCN-MCD+DE	86.2 $\pm$ 1.7/72.6 $\pm$ 0.1	87.3 $\pm$ 0.1	83.9 $\pm$ 1.5/66.6 $\pm$ 0.2	83.0 $\pm$ 0.2	64.7 $\pm$ 1.5/53.7 $\pm$ 0.2	72.1 $\pm$ 0.2	47.0 $\pm$ 1.4/68.8 $\pm$ 0.2	72.1 $\pm$ 0.2	27.8 $\pm$ 1.5/59.0 $\pm$ 0.2	72.1 $\pm$ 0.2	66.9 $\pm$ 1.5/69.7 $\pm$ 0.2	78.9 $\pm$ 0.2	70.9 $\pm$ 1.5/64.4 $\pm$ 0.2	85.9 $\pm$ 0.2
	GCN-BNN	87.5 $\pm$ 1.7/96.0 $\pm$ 0.1	89.0 $\pm$ 0.1	83.8 $\pm$ 1.5/82.7 $\pm$ 0.2	84.9 $\pm$ 0.2	64.6 $\pm$ 1.5/60.9 $\pm$ 0.2	73.1 $\pm$ 0.2	49.9 $\pm$ 1.5/70.7 $\pm$ 0.2	73.1 $\pm$ 0.2	32.6 $\pm$ 1.5/56.4 $\pm$ 0.2	73.1 $\pm$ 0.2	69.1 $\pm$ 1.5/69.1 $\pm$ 0.2	78.9 $\pm$ 0.2	71.8 $\pm$ 1.5/67.0 $\pm$ 0.2	88.6 $\pm$ 0.2
	GCN-Ens	89.8 $\pm$ 1.7/99.9 $\pm$ 0.1	90.4 $\pm$ 0.1	87.6 $\pm$ 1.5/88.4 $\pm$ 0.2	87.3 $\pm$ 0.2	66.6 $\pm$ 1.5/60.5 $\pm$ 0.2	75.9 $\pm$ 0.2	59.9 $\pm$ 1.5/58.6 $\pm$ 0.2	75.9 $\pm$ 0.2	31.7 $\pm$ 1.5/55.7 $\pm$ 0.2	75.9 $\pm$ 0.2	73.0 $\pm$ 1.5/69.2 $\pm$ 0.2	82.6 $\pm$ 0.2	71.3 $\pm$ 1.5/67.1 $\pm$ 0.2	91.5 $\pm$ 0.2
	GCN	85.3 $\pm$ 1.8/88.1 $\pm$ 0.1	88.5 $\pm$ 0.1	85.1 $\pm$ 1.5/82.9 $\pm$ 0.2	86.4 $\pm$ 0.2	54.1 $\pm$ 1.5/53.4 $\pm$ 0.2	72.8 $\pm$ 0.2	53.4 $\pm$ 1.5/52.4 $\pm$ 0.2	72.8 $\pm$ 0.2	51.4 $\pm$ 1.5/55.5 $\pm$ 0.2	72.8 $\pm$ 0.2	52.4 $\pm$ 1.5/67.1 $\pm$ 0.2	78.4 $\pm$ 0.2	66.9 $\pm$ 1.5/61.5 $\pm$ 0.2	87.6 $\pm$ 0.2
	SGCN	89.2 $\pm$ 1.7/98.1 $\pm$ 0.2	90.2 $\pm$ 0.2	89.8 $\pm$ 1.5/88.9 $\pm$ 0.2	89.9 $\pm$ 0.2	54.9 $\pm$ 1.5/52.9 $\pm$ 0.2	76.2 $\pm$ 0.2	39.1 $\pm$ 1.5/69.7 $\pm$ 0.2	76.2 $\pm$ 0.2	29.1 $\pm$ 1.5/45.0 $\pm$ 0.2	76.2 $\pm$ 0.2	74.1 $\pm$ 1.5/57.9 $\pm$ 0.2	78.9 $\pm$ 0.2	72.9 $\pm$ 1.5/75.1 $\pm$ 0.2	90.0 $\pm$ 0.2
	GCN-EBM	89.7 $\pm$ 1.7/99.9 $\pm$ 0.1	90.3 $\pm$ 0.1	87.6 $\pm$ 1.5/88.4 $\pm$ 0.2	87.3 $\pm$ 0.2	67.2 $\pm$ 1.5/67.3 $\pm$ 0.2	75.8 $\pm$ 0.2	59.9 $\pm$ 1.5/58.6 $\pm$ 0.2	75.8 $\pm$ 0.2	31.7 $\pm$ 1.5/55.7 $\pm$ 0.2	75.8 $\pm$ 0.2	73.0 $\pm$ 1.5/69.2 $\pm$ 0.2	82.6 $\pm$ 0.2	71.3 $\pm$ 1.5/67.1 $\pm$ 0.2	91.2 $\pm$ 0.2
	GCN-HEAT	89.7 $\pm$ 1.7/98.1 $\pm$ 0.2	90.3 $\pm$ 0.2	87.6 $\pm$ 1.5/88.2 $\pm$ 0.2	87.4 $\pm$ 0.2	67.2 $\pm$ 1.5/65.5 $\pm$ 0.2	75.8 $\pm$ 0.2	59.9 $\pm$ 1.5/58.6 $\pm$ 0.2	75.8 $\pm$ 0.2	31.7 $\pm$ 1.5/55.7 $\pm$ 0.2	75.8 $\pm$ 0.2	73.0 $\pm$ 1.5/69.2 $\pm$ 0.2	82.6 $\pm$ 0.2	71.3 $\pm$ 1.5/67.1 $\pm$ 0.2	91.2 $\pm$ 0.2
	GCN-Safe	89.7 $\pm$ 1.7/98.1 $\pm$ 0.2	90.3 $\pm$ 0.2	87.6 $\pm$ 1.5/90.5 $\pm$ 0.2	87.4 $\pm$ 0.2	67.2 $\pm$ 1.5/67.0 $\pm$ 0.2	75.8 $\pm$ 0.2	59.9 $\pm$ 1.5/58.6 $\pm$ 0.2	75.8 $\pm$ 0.2	31.7 $\pm$ 1.5/55.7 $\pm$ 0.2	75.8 $\pm$ 0.2	73.0 $\pm$ 1.5/69.2 $\pm$ 0.2	82.6 $\pm$ 0.2	71.3 $\pm$ 1.5/67.1 $\pm$ 0.2	91.2 $\pm$ 0.2
	GCN-GERM	89.7 $\pm$ 1.7/98.1 $\pm$ 0.2	90.3 $\pm$ 0.2	87.6 $\pm$ 1.5/89.9 $\pm$ 0.2	87.4 $\pm$ 0.2	67.2 $\pm$ 1.5/67.7 $\pm$ 0.2	75.8 $\pm$ 0.2	59.9 $\pm$ 1.5/58.6 $\pm$ 0.2	75.8 $\pm$ 0.2	31.7 $\pm$ 1.5/55.7 $\pm$ 0.2	75.8 $\pm$ 0.2	73.0 $\pm$ 1.5/69.2 $\pm$ 0.2	82.6 $\pm$ 0.2	71.3 $\pm$ 1.5/67.1 $\pm$ 0.2	91.2 $\pm$ 0.2
CoNLL-LLM	GCN-MCD	87.1 $\pm$ 1.5/49.9 $\pm$ 0.2	90.1 $\pm$ 0.2	86.6 $\pm$ 1.5/40.1 $\pm$ 0.2	89.5 $\pm$ 0.2	n.a./n.a.	n.a.	70.1 $\pm$ 1.5/50.0 $\pm$ 0.2	76.9 $\pm$ 0.2	37.8 $\pm$ 1.5/49.7 $\pm$ 0.2	76.9 $\pm$ 0.2	70.5 $\pm$ 1.5/50.1 $\pm$ 0.2	84.0 $\pm$ 0.2	70.2 $\pm$ 1.5/50.2 $\pm$ 0.2	91.0 $\pm$ 0.2
	GCN-DE	84.8 $\pm$ 1.7/70.9 $\pm$ 0.2	87.8 $\pm$ 0.2	85.2 $\pm$ 1.5/70.2 $\pm$ 0.2	87.9 $\pm$ 0.2	n.a./n.a.	n.a.	74.7 $\pm$ 1.5/52.7 $\pm$ 0.2	75.6 $\pm$ 0.2	35.9 $\pm$ 1.5/63.2 $\pm$ 0.2	75.6 $\pm$ 0.2	64.4 $\pm$ 1.5/69.2 $\pm$ 0.2	81.8 $\pm$ 0.2	68.7 $\pm$ 1.5/68.1 $\pm$ 0.2	88.5 $\pm$ 0.2
	GCN-MCD+DE	84.5 $\pm$ 1.7/71.3 $\pm$ 0.2	88.1 $\pm$ 0.2	85.2 $\pm$ 1.5/70.7 $\pm$ 0.2	87.9 $\pm$ 0.2	n.a./n.a.	n.a.	75.1 $\pm$ 1.5/52.7 $\pm$ 0.2	75.2 $\pm$ 0.2	35.4 $\pm$ 1.5/63.7 $\pm$ 0.2	75.2 $\pm$ 0.2	64.7 $\pm$ 1.5/70.1 $\pm$ 0.2	81.3 $\pm$ 0.2	68.7 $\pm$ 1.5/68.9 $\pm$ 0.2	88.7 $\pm$ 0.2
	GCN-BNN	86.3 $\pm$ 1.7/90.2 $\pm$ 0.2	90.0 $\pm$ 0.2	86.2 $\pm$ 1.5/81.8 $\pm$ 0.2	89.3 $\pm$ 0.2	n.a./n.a.	n.a.	71.1 $\pm$ 1.5/57.3 $\pm$ 0.2	76.8 $\pm$ 0.2	39.9 $\pm$ 1.5/67.6 $\pm$ 0.2	76.8 $\pm$ 0.2	68.4 $\pm$ 1.5/69.0 $\pm$ 0.2	83.7 $\pm$ 0.2	71.6 $\pm$ 1.5/72.8 $\pm$ 0.2	90.3 $\pm$ 0.2
	GCN-Ens	87.1 $\pm$ 1.5/79.5 $\pm$ 0.2	90.1 $\pm$ 0.2	86.7 $\pm$ 1.5/76.3 $\pm$ 0.2	89.5 $\pm$ 0.2	n.a./n.a.	n.a.	70.8 $\pm$ 1.5/46.5 $\pm$ 0.2	77.3 $\pm$ 0.2	35.8 $\pm$ 1.5/68.4 $\pm$ 0.2	77.3 $\pm$ 0.2	70.4 $\pm$ 1.5/63.2 $\pm$ 0.2	84.2 $\pm$ 0.2	70.2 $\pm$ 1.5/69.9 $\pm$ 0.2	91.3 $\pm$ 0.2
	GCN	83.6 $\pm$ 1.8/88.2 $\pm$ 0.1	89.4 $\pm$ 0.1	86.2 $\pm$ 1.5/86.7 $\pm$ 0.2	88.1 $\pm$ 0.2	n.a./n.a.	n.a.	56.4 $\pm$ 1.5/57.2 $\pm$ 0.2	75.8 $\pm$ 0.2	53.7 $\pm$ 1.5/58.6 $\pm$ 0.2	75.8 $\pm$ 0.2	52.6 $\pm$ 1.5/66.3 $\pm$ 0.2	81.3 $\pm$ 0.2	68.0 $\pm$ 1.5/63.4 $\pm$ 0.2	90.3 $\pm$ 0.2
	SGCN	87.4 $\pm$ 1.7/90.0 $\pm$ 0.2	89.9 $\pm$ 0.2	86.2 $\pm$ 1.5/85.0 $\pm$ 0.2	89.1 $\pm$ 0.2	n.a./n.a.	n.a.	60.3 $\pm$ 1.5/66.3 $\pm$ 0.2	68.1 $\pm$ 0.2	36.8 $\pm$ 1.5/40.4 $\pm$ 0.2	68.1 $\pm$ 0.2	69.5 $\pm$ 1.5/65.4 $\pm$ 0.2	78.9 $\pm$ 0.2	64.9 $\pm$ 1.5/66.1 $\pm$ 0.2	80.2 $\pm$ 0.2
	GCN-EBM	86.9 $\pm$ 1.7/87.1 $\pm$ 0.2	90.2 $\pm$ 0.2	86.7 $\pm$ 1.5/87.5 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a./n.a.	n.a.	70.8 $\pm$ 1.5/70.6 $\pm$ 0.2	77.1 $\pm$ 0.2	35.8 $\pm$ 1.5/33.0 $\pm$ 0.2	77.1 $\pm$ 0.2	70.2 $\pm$ 1.5/72.7 $\pm$ 0.2	83.7 $\pm$ 0.2	70.1 $\pm$ 1.5/69.9 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-HEAT	86.9 $\pm$ 1.7/88.4 $\pm$ 0.2	90.2 $\pm$ 0.2	86.7 $\pm$ 1.5/88.4 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a./n.a.	n.a.	70.8 $\pm$ 1.5/71.9 $\pm$ 0.2	77.1 $\pm$ 0.2	35.8 $\pm$ 1.5/35.9 $\pm$ 0.2	77.1 $\pm$ 0.2	70.2 $\pm$ 1.5/72.1 $\pm$ 0.2	83.7 $\pm$ 0.2	70.1 $\pm$ 1.5/69.8 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-Safe	86.9 $\pm$ 1.7/88.4 $\pm$ 0.2	90.2 $\pm$ 0.2	86.7 $\pm$ 1.5/91.0 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a./n.a.	n.a.	70.8 $\pm$ 1.5/70.6 $\pm$ 0.2	77.1 $\pm$ 0.2	35.8 $\pm$ 1.5/42.0 $\pm$ 0.2	77.1 $\pm$ 0.2	70.2 $\pm$ 1.5/70.7 $\pm$ 0.2	83.7 $\pm$ 0.2	70.1 $\pm$ 1.5/71.3 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-GERM	86.9 $\pm$ 1.7/88.4 $\pm$ 0.2	90.2 $\pm$ 0.2	86.7 $\pm$ 1.5/89.7 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a./n.a.	n.a.	70.8 $\pm$ 1.5/82.2 $\pm$ 0.2	77.1 $\pm$ 0.2	35.8 $\pm$ 1.5/85.3 $\pm$ 0.2	77.1 $\pm$ 0.2	70.2 $\pm$ 1.5/91.9 $\pm$ 0.2	83.7 $\pm$ 0.2	70.1 $\pm$ 1.5/75.0 $\pm$ 0.2	91.1 $\pm$ 0.2
Chesnut	GCN-MCD	86.7 $\pm$ 1.5/50.7 $\pm$ 0.2	84.8 $\pm$ 0.2	88.3 $\pm$ 1.5/50.1 $\pm$ 0.2	90.9 $\pm$ 0.2	66.9 $\pm$ 1.5/49.1 $\pm$ 0.2	76.4 $\pm$ 0.2	53.0 $\pm$ 1.5/49.5 $\pm$ 0.2	76.4 $\pm$ 0.2	29.4 $\pm$ 1.5/50.3 $\pm$ 0.2	76.4 $\pm$ 0.2	65.6 $\pm$ 1.5/49.2 $\pm$ 0.2	78.4 $\pm$ 0.2	51.9 $\pm$ 1.5/50.2 $\pm$ 0.2	77.3 $\pm$ 0.2
	GCN-DE	81.0 $\pm$ 1.5/60.8 $\pm$ 0.2	81.0 $\pm$ 0.2	83.1 $\pm$ 1.5/69.1 $\pm$ 0.2	86.4 $\pm$ 0.2	62.9 $\pm$ 1.5/56.6 $\pm$ 0.2	72.9 $\pm$ 0.2	50.1 $\pm$ 1.5/79.3 $\pm$ 0.2	72.9 $\pm$ 0.2	27.3 $\pm$ 1.5/57.2 $\pm$ 0.2	72.9 $\pm$ 0.2	65.6 $\pm$ 1.5/62.0 $\pm$ 0.2	75.6 $\pm$ 0.2	51.9 $\pm$ 1.5/49.9 $\pm$ 0.2	74.3 $\pm$ 0.2
	GCN-MCD+DE	81.8 $\pm$ 1.7/70.8 $\pm$ 0.2	81.3 $\pm$ 0.2	84.0 $\pm$ 1.5/69.9 $\pm$ 0.2	87.1 $\pm$ 0.2	62.7 $\pm$ 1.5/57.5 $\pm$ 0.2	73.0 $\pm$ 0.2	50.3 $\pm$ 1.5/79.6 $\pm$ 0.2	73.0 $\pm$ 0.2	24.4 $\pm$ 1.5/62.4 $\pm$ 0.2	73.0 $\pm$ 0.2	59.3 $\pm$ 1.5/59.9 $\pm$ 0.2	75.0 $\pm$ 0.2	51.5 $\pm$ 1.5/60.2 $\pm$ 0.2	74.6 $\pm$ 0.2
	GCN-BNN	86.3 $\pm$ 1.7/96.0 $\pm$ 0.1	84.5 $\pm$ 0.2	87.3 $\pm$ 1.5/86.6 $\pm$ 0.2	90.3 $\pm$ 0.2	64.7 $\pm$ 1.5/54.3 $\pm$ 0.2	75.8 $\pm$ 0.2	43.7 $\pm$ 1.5/72.7 $\pm$ 0.2	75.8 $\pm$ 0.2	27.6 $\pm$ 1.5/61.2 $\pm$ 0.2	75.8 $\pm$ 0.2	63.2 $\pm$ 1.5/55.9 $\pm$ 0.2	77.8 $\pm$ 0.2	52.8 $\pm$ 1.5/64.1 $\pm$ 0.2	77.1 $\pm$ 0.2
	GCN-Ens	87.0 $\pm$ 1.5/75.5 $\pm$ 0.2	84.8 $\pm$ 0.2	86.7 $\pm$ 1.5/77.7 $\pm$ 0.2	91.1 $\pm$ 0.2	65.6 $\pm$ 1.5/54.8 $\pm$ 0.2	76.9 $\pm$ 0.2	49.9 $\pm$ 1.5/69.7 $\pm$ 0.2	76.9 $\pm$ 0.2	29.1 $\pm$ 1.5/57.7 $\pm$ 0.2	76.9 $\pm$ 0.2	63.6 $\pm$ 1.5/68.0 $\pm$ 0.2	78.6 $\pm$ 0.2	53.9 $\pm$ 1.5/55.4 $\pm$ 0.2	77.7 $\pm$ 0.2
	GCN	88.0 $\pm$ 1.8/88.4 $\pm$ 0.2	85.6 $\pm$ 0.2	90.1 $\pm$ 1.5/84.3 $\pm$ 0.2	91.6 $\pm$ 0.2	58.1 $\pm$ 1.5/54.4 $\pm$ 0.2	76.7 $\pm$ 0.2	43.9 $\pm$ 1.5/66.6 $\pm$ 0.2	76.7 $\pm$ 0.2	30.9 $\pm$ 1.5/61.9 $\pm$ 0.2	76.7 $\pm$ 0.2	67.7 $\pm$ 1.5/65.1 $\pm$ 0.2	78.3 $\pm$ 0.2	53.9 $\pm$ 1.5/60.3 $\pm$ 0.2	77.4 $\pm$ 0.2
	SGCN	86.1 $\pm$ 1.8/84.9 $\pm$ 0.2	85.0 $\pm$ 0.2	88.0 $\pm$ 1.5/85.4 $\pm$ 0.2	90.4 $\pm$ 0.2	64.5 $\pm$ 1.5/61.8 $\pm$ 0.2	75.8 $\pm$ 0.2	63.6 $\pm$ 1.5/64.1 $\pm$ 0.2	75.8 $\pm$ 0.2	28.1 $\pm$ 1.5/25.6 $\pm$ 0.2	75.8 $\pm$ 0.2	63.7 $\pm$ 1.5/60.8 $\pm$ 0.2	79.3 $\pm$ 0.2	52.1 $\pm$ 1.5/62.2 $\pm$ 0.2	76.0 $\pm$ 0.2
	GCN-EBM	86.8 $\pm$ 1.7/87.4 $\pm$ 0.2	84.8 $\pm$ 0.2	88.3 $\pm$ 1.5/88.0 $\pm$ 0.2	90.9 $\pm$ 0.2	66.2 $\pm$ 1.5/64.7 $\pm$ 0.2	76.6 $\pm$ 0.2	55.1 $\pm$ 1.5/43.8 $\pm$ 0.2	76.6 $\pm$ 0.2	29.6 $\pm$ 1.5/23.5 $\pm$ 0.2	76.6 $\pm$ 0.2	65.7 $\pm$ 1.5/66.4 $\pm$ 0.2	78.7 $\pm$ 0.		

Model	CoraML	CoraML LLM	Citeseer	PubMed	Amazon Computers	Amazon Photo	Coauthor CS	Coauthor Physics
GCN-MCD	10.7/20.1	10.1/19.9	9.8/19.3	9.2/18.8	10.2/19.6	10.2/19.7	10.0/19.3	10.5/20.4
GCN-DE	6.3/14.2	8.0/16.9	7.8/16.2	7.3/13.6	7.2/15.9	7.1/15.8	6.5/13.2	8.1/16.8
GCN-MCD+DE	6.9/14.8	7.2/15.9	6.6/14.7	7.3/13.6	7.7/16.3	7.4/16.2	6.8/13.5	7.8/16.4
GCN-BNN	5.2/11.4	6.0/13.4	5.4/11.0	5.3/10.9	4.3/9.0	5.1/9.8	4.9/10.5	4.8/9.9
GCN-Ens	7.8/16.1	6.3/15.0	6.4/12.9	6.2/12.4	5.7/11.6	4.7/9.0	4.8/10.4	6.2/13.3
GNP	7.0/15.2	5.4/11.2	6.6/12.9	5.8/12.4	5.7/11.8	5.3/11.7	7.2/13.9	6.1/14.5
SGCN	5.2/9.5	6.1/14.6	6.0/12.6	5.0/11.7	9.1/18.5	7.6/15.1	9.3/17.9	4.6/11.2
GCN-EBM	4.5/7.9	4.0/8.2	5.2/10.3	5.2/12.0	4.2/10.7	4.9/10.9	4.1/7.1	4.4/8.4
GCN-HEAT	4.4/10.4	5.4/12.9	3.8/9.2	5.6/12.8	3.5/8.6	4.5/10.2	4.2/8.3	4.7/10.4
GCNSafe	5.3/9.1	4.4/8.1	5.2/9.5	5.4/8.9	5.9/13.2	6.4/12.9	5.5/10.7	5.8/11.7
GCN-GEBM	2.7 / 4.5	2.9 / 4.9	3.3 / 5.7	3.8 / 7.4	2.5 / 4.2	2.7 / 4.5	2.7 / 4.3	3.0 / 5.0

Table 7: Average o.o.d. detection rank (AUC-ROC) ($\downarrow$) of epistemic uncertainty versus other epistemic measures / all uncertainty measures over all distribution shifts in an inductive setting (**best** and **runner-up**).

for the GCN used at the backbone of our experiments. Note that since our model does not change the output of the classifier, the calibration could be further improved using post hoc methods like temperature scaling [26]. This, however, is beyond the scope of this work.

Expected Calibration Error (ECE). The expected calibration error ($\downarrow$) [50] measures how well the predicted probabilities that are normalized to the interval $[0, 1]$ match the true predictive accuracy. To that end, we bin each prediction into $B = 20$ bins according to their confidence (i.e. maximum softmax response $\max_c p(y = c | \mathbf{x})$). For each bin, we then compute the average accuracy and compute the ECE as the mean over all bins weighted by their size.

$$\text{ECE} = \sum_k \frac{|B_k|}{n} |\text{accuracy}(B_k) - \text{confidence}(B_k)|$$

Brier Score. A similar metric is the Brier score [7] ($\downarrow$) which computes the mean squared distance between the predicted probabilities and the one-hot encoded true labels.

$$\text{Brier} = \frac{1}{n} \sum_i \|p(\mathbf{y} | \mathbf{x}_i) - \mathbf{y}_i\|_2^2$$

We report ECE and Brier scores for inductive and transductive settings in Tables 20 to 23. In terms of calibration metrics, none of the considered approaches consistently shows strong merits. We remark that work on GEBM and GNN calibration is somewhat orthogonal and our framework can be applied to any well-calibrated GNN backbone.

C.5 Backbone Architecture

We report AUC-ROC and AUC-PR for o.o.d. detection using GEBM and different GNN backbones with corresponding standard deviations in Tables 24 and 25. GEBM is effective on all models and achieves the highest scores on most distribution shifts.

C.6 Bias of Diffusion Operators

At the core of many GNNs and also our GEBM framework lies a diffusion operator $P_A : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Here, we discuss three typical realizations and the bias they introduce.

Symmetric Diffusion. Symmetric diffusion normalizes the adjacency matrix as $\hat{A} = D^{-1/2} A D^{-1/2}$. Here, $D = \text{diag}(\deg(v_1), \dots, \deg(v_n))$ is a diagonal matrix of node degrees. The symmetric diffusion operator has a dominant eigenvector that correlates with the node degree $\mathbf{v}_{\max} \propto \mathbf{d}^{1/2}$ and therefore also centrality [12]. Repeated application of this diffusion process to any arbitrary signal will concentrate confidence at high degree nodes. Applying this diffusion to a confidence measure like the logits of GNN will therefore always favor high degree nodes.

	Model	LoC (test)		LoC (train)		Bert(p) (train)		Bert(0.5)		A/O(1) (test)		Page Rank		Homophily			
		AUC-ROC (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†	AUC-ROC ^(p) (Aleat./Epist.)	Acc.†		
ConcML	GCN-MCD	87.5 _{±1.0}	90.3 _{±0.7}	90.4 _{±0.8}	84.9 _{±1.0}	79.0 _{±1.2}	85.2 _{±1.3}	67.5 _{±1.1}	79.0 _{±1.3}	82.0 _{±1.4}	61.8 _{±1.0}	79.0 _{±1.3}	82.0 _{±1.4}	71.4 _{±1.0}	82.6 _{±1.3}	69.0 _{±1.0}	89.3 _{±1.1}
	GCN-DE	84.6 _{±1.7}	71.9 _{±1.3}	87.3 _{±1.3}	81.9 _{±1.0}	67.7 _{±1.1}	75.7 _{±1.3}	78.4 _{±1.3}	52.2 _{±1.0}	77.0 _{±1.3}	78.4 _{±1.3}	52.2 _{±1.0}	77.0 _{±1.3}	27.2 _{±1.1}	55.6 _{±1.3}	78.4 _{±1.3}	64.3 _{±1.0}
	GCN-MCD+DE	84.6 _{±1.7}	72.2 _{±1.3}	87.6 _{±1.3}	80.9 _{±1.0}	66.8 _{±1.1}	75.8 _{±1.3}	78.1 _{±1.3}	52.9 _{±1.0}	77.1 _{±1.3}	78.1 _{±1.3}	52.9 _{±1.0}	77.1 _{±1.3}	27.9 _{±1.1}	55.7 _{±1.3}	78.1 _{±1.3}	64.7 _{±1.0}
	GCN-BNN	84.7 _{±1.7}	78.1 _{±1.3}	90.0 _{±0.8}	78.6 _{±1.0}	67.3 _{±1.1}	77.3 _{±1.3}	80.6 _{±1.3}	53.5 _{±1.0}	77.0 _{±1.3}	80.6 _{±1.3}	53.5 _{±1.0}	77.0 _{±1.3}	32.8 _{±1.1}	75.4 _{±1.3}	80.6 _{±1.3}	67.6 _{±1.0}
	GCN-Ens	87.6 _{±1.7}	77.6 _{±1.3}	93.0 _{±0.7}	84.9 _{±1.0}	67.7 _{±1.1}	75.1 _{±1.3}	81.8 _{±1.3}	61.4 _{±1.0}	76.4 _{±1.3}	81.8 _{±1.3}	61.4 _{±1.0}	76.4 _{±1.3}	32.3 _{±1.1}	75.9 _{±1.3}	81.8 _{±1.3}	71.5 _{±1.0}
	GCN-ERM	84.7 _{±1.7}	78.8 _{±1.3}	90.0 _{±0.8}	82.2 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	SGCN	87.7 _{±1.7}	78.0 _{±1.3}	91.0 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-EBM	87.6 _{±1.7}	78.0 _{±1.3}	91.0 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-HEAT	87.6 _{±1.7}	78.0 _{±1.3}	91.0 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-Safe	87.6 _{±1.7}	78.0 _{±1.3}	91.0 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
ConcML-LM	GCN-MCD	88.3 _{±1.7}	75.0 _{±1.3}	91.2 _{±0.7}	86.5 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-DE	80.8 _{±1.7}	77.1 _{±1.3}	89.7 _{±1.3}	85.1 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-MCD+DE	87.0 _{±1.7}	77.5 _{±1.3}	89.4 _{±0.8}	85.0 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-BNN	88.2 _{±1.7}	78.0 _{±1.3}	91.4 _{±0.7}	86.0 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-Ens	88.3 _{±1.7}	78.1 _{±1.3}	91.4 _{±0.7}	86.0 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-ERM	86.3 _{±1.7}	79.0 _{±1.3}	90.5 _{±0.7}	85.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	SGCN	89.3 _{±1.7}	79.0 _{±1.3}	90.5 _{±0.7}	87.7 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-EBM	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-HEAT	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-Safe	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
Chisner	GCN-MCD	88.2 _{±1.7}	78.0 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-DE	81.0 _{±1.7}	77.0 _{±1.3}	90.2 _{±0.7}	81.1 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-MCD+DE	81.7 _{±1.7}	77.0 _{±1.3}	90.2 _{±0.7}	81.1 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-BNN	83.8 _{±1.7}	78.3 _{±1.3}	92.5 _{±0.7}	83.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-Ens	85.4 _{±1.7}	77.3 _{±1.3}	93.8 _{±0.7}	86.1 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-ERM	84.2 _{±1.7}	78.9 _{±1.3}	91.4 _{±0.7}	87.8 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	SGCN	89.3 _{±1.7}	79.0 _{±1.3}	90.5 _{±0.7}	87.7 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-EBM	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-HEAT	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
	GCN-Safe	88.2 _{±1.7}	78.2 _{±1.3}	91.4 _{±0.7}	86.6 _{±1.0}	67.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	80.6 _{±1.3}	52.9 _{±1.0}	77.3 _{±1.3}	32.3 _{±1.1}	75.3 _{±1.3}	80.6 _{±1.3}	67.3 _{±1.0}
DabMed	GCN-MCD	66.0 _{±1.1}	50.1 _{±1.3}	91.6 _{±1.4}	63.2 _{±1.0}	50.2 _{±1.3}	83.7 _{±1.3}	n.a./n.a.	n.a.	69.3 _{±1.1}	49.1 _{±1.3}	82.1 _{±1.3}	62.1 _{±1.0}	50.1 _{±1.3}	82.9 _{±1.3}	51.0 _{±1.0}	51.0 _{±1.0}
	GCN-DE	65.1 _{±1.1}	57.2 _{±1.3}	91.2 _{±0.7}	63.8 _{±1.0}	56.8 _{±1.3}	80.5 _{±1.3}	n.a./n.a.	n.a.	62.6 _{±1.0}	50.1 _{±1.3}	74.9 _{±1.3}	60.5 _{±1.0}	57.0 _{±1.3}	74.9 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	GCN-MCD+DE	65.8 _{±1.1}	57.5 _{±1.3}	90.8 _{±0.7}	64.1 _{±1.0}	56.9 _{±1.3}	80.7 _{±1.3}	n.a./n.a.	n.a.	62.1 _{±1.0}	50.1 _{±1.3}	75.4 _{±1.3}	60.5 _{±1.0}	57.0 _{±1.3}	75.4 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	GCN-BNN	66.0 _{±1.1}	56.1 _{±1.3}	90.7 _{±1.4}	63.1 _{±1.0}	56.3 _{±1.3}	84.0 _{±1.3}	n.a./n.a.	n.a.	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	GCN-Ens	66.0 _{±1.1}	56.1 _{±1.3}	90.7 _{±1.4}	63.1 _{±1.0}	56.3 _{±1.3}	84.0 _{±1.3}	n.a./n.a.	n.a.	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	GCN-ERM	66.1 _{±1.1}	56.2 _{±1.3}	91.1 _{±0.7}	63.2 _{±1.0}	56.4 _{±1.3}	83.8 _{±1.3}	n.a./n.a.	n.a.	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	SGCN	66.1 _{±1.1}	56.2 _{±1.3}	91.1 _{±0.7}	63.2 _{±1.0}	56.4 _{±1.3}	83.8 _{±1.3}	n.a./n.a.	n.a.	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	60.1 _{±1.0}	50.1 _{±1.3}	78.2 _{±1.3}	54.7 _{±1.0}	54.7 _{±1.0}
	GCN-EBM	66.1 _{±1.1}	56.2 _{±1.3}	91.1 _{±0.7}	63.2 _{±1.0}	56.4 _{±1.3}	83.8 _{±1.3}	n.a./n.a.	n.a.	60.1<							

Model	CoraML	CoraML LLM	Citeseer	PubMed	Amazon Computers	Amazon Photo	Coauthor CS	Coauthor Physics
GCN-MCD	10.5/20.3	10.2/20.0	10.2/19.8	8.9/18.5	10.4/19.9	10.7/20.2	10.2/19.3	10.7/20.6
GCN-DE	6.7/14.3	8.0/16.9	7.5/15.6	7.2/13.5	6.6/15.4	7.3/15.8	6.6/14.3	7.8/16.6
GCN-MCD+DE	6.8/14.5	7.6/16.5	7.2/15.2	7.0/13.3	6.5/15.3	6.6/14.7	6.8/13.7	7.9/16.7
GCN-BNN	5.7/11.9	6.1/13.7	5.3/11.3	5.1/11.4	4.7/11.9	3.8/7.9	5.1/10.8	4.9/11.6
GCN-Ens	7.4/14.1	5.7/14.1	5.8/12.4	7.7/14.2	7.3/15.9	5.3/11.5	4.3/9.5	5.7/12.8
GPV	6.5/12.8	6.1/10.9	7.5/15.7	5.8/11.3	4.8/10.8	5.1/11.1	6.7/13.1	6.2/13.0
SGCN	3.9/6.3	4.4/7.5	4.3/7.9	3.9/8.2	8.1/17.2	6.7/14.0	9.2/17.7	4.4/9.9
GCN-EBM	5.2/9.7	5.1/10.7	5.1/11.4	5.9/13.4	4.6/12.5	5.6/13.3	4.3/8.3	4.6/9.4
GCN-HEAT	4.7/9.3	4.8/12.2	3.8/9.2	5.1/12.2	3.9/9.9	5.6/14.1	4.4/9.2	4.9/11.1
GCNSafe	5.8/9.6	4.4/9.4	5.4/9.9	5.6/9.1	6.5/13.8	6.5/13.4	5.7/11.0	5.9/11.6
GCN-GEBM	2.8 / 4.7	3.4 / 5.9	3.9 / 6.9	3.8 / 5.8	2.7 / 6.0	3.0 / 5.2	2.7 / 4.7	3.0 / 5.0

Table 9: Average o.o.d. detection rank (AUC-ROC) ($\downarrow$) of epistemic uncertainty versus other epistemic measures / all uncertainty measures over all distribution shifts in a transductive setting (**best** and **runner-up**).

marginalization, as discussed in Section 4.3. We want to highlight that GEBM allows to incorporate additional energy functions as well, e.g. if such information is available in designated downstream applications. As shown in this study, we find that the proposed energy functions already suffice to enable GEBM to be highly sensitive to various distribution shifts.

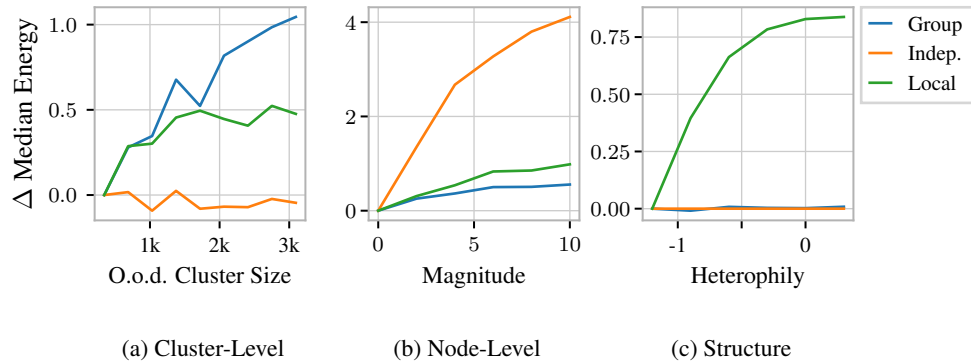


Figure 4: Energy of different types for anomalies of increasing severity on synthetic data. We vary the size of an o.o.d. cluster on real data (left), insert per-node anomalies to the energies of an SBM (middle), and increase the heterophily in an SBM (right).

Node-Level Anomalies. We study node level anomalies by generating synthetic CSBM graphs [57] (intra-edge probability 0.05, inter-edge probability 0.001, $n = 1000$, $c = 7$). We then generate c -dimensional logits (intuitively confidence scores) according to a standard normal distribution that is centered at $l_{\text{disp}} = 4$ for the dimension of the true class label of a node and 0 otherwise. This roughly resembles the logits of a well-trained classifier which peak at the true unknown class. We then perturb a fraction $p = 0.05$ of the logits by replacing them with positive noise from a standard normal distribution that is rescaled by an increasing magnitude. For each magnitude, we compute all three energy terms for each node and compare the difference in energy to the corresponding unperturbed energy. Effectively, this monitors how the energy of a node changes under increasingly severe distribution shifts that affect the logits of a classifier. We visualize the median of the energy differences in Figure 4b: While all energies increase with higher magnitudes, the structure-agnostic energy term of GEBM is the most sensitive to anomalies that are located at individual nodes. It does not suffer from the smoothing of anomalous scores induced by graph diffusion.

Cluster-Level Anomalies. We study anomalies that affect an entire cluster of nodes by iteratively introducing an anomalous cluster to the graph. To that end, we leave out k classes and train a GCN on the Amazon Photos dataset. At inference, we re-introduce nodes (and edges) of the left-out classes which can be seen as anomalous from the perspective of the classifier. Similar to the aforementioned node-level anomaly, we monitor how each energy type changes while iteratively introducing the entire cluster(s) of o.o.d. nodes into the graph. Again, we see that the cluster-level energy term of GEBM is the most sensitive to this distribution shift in Figure 4a.

	Model	LoC (<i>last</i>)		LoC (<i>hetero</i>)		Ber(β) (<i>neur</i>)		Ber(0.5)		N(0,1) (<i>fir</i>)		Page Rank		Homophily	
		AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$	AUC-PR $\uparrow$ (Area Fpr)	Acc. $\uparrow$
CoraML	GCN-MCD	85.1 $\pm$ 0.466.5 $\pm$ 0.2	90.4 $\pm$ 0.2	74.0 $\pm$ 0.433.4 $\pm$ 0.2	87.4 $\pm$ 0.2	64.6 $\pm$ 0.551.4 $\pm$ 0.2	75.6 $\pm$ 0.2	54.2 $\pm$ 0.551.0 $\pm$ 0.2	75.6 $\pm$ 0.2	37.8 $\pm$ 0.447.6 $\pm$ 0.2	75.6 $\pm$ 0.2	71.8 $\pm$ 0.566.7 $\pm$ 0.2	82.3 $\pm$ 0.2	66.1 $\pm$ 0.560.8 $\pm$ 0.2	91.4 $\pm$ 0.2
	GCN-DE	81.0 $\pm$ 0.467.9 $\pm$ 0.2	87.2 $\pm$ 0.2	66.8 $\pm$ 0.412.2 $\pm$ 0.2	82.0 $\pm$ 0.2	61.7 $\pm$ 0.541.3 $\pm$ 0.2	71.9 $\pm$ 0.2	45.9 $\pm$ 0.609.9 $\pm$ 0.2	71.9 $\pm$ 0.2	36.0 $\pm$ 0.461.5 $\pm$ 0.2	71.9 $\pm$ 0.2	66.3 $\pm$ 0.468.4 $\pm$ 0.2	79.3 $\pm$ 0.2	66.3 $\pm$ 0.468.4 $\pm$ 0.2	85.5 $\pm$ 0.2
	GCN-MCD+DE	81.3 $\pm$ 0.462.3 $\pm$ 0.2	87.3 $\pm$ 0.2	69.0 $\pm$ 0.405.4 $\pm$ 0.2	83.0 $\pm$ 0.2	61.8 $\pm$ 0.533.7 $\pm$ 0.2	72.1 $\pm$ 0.2	45.5 $\pm$ 0.607.4 $\pm$ 0.2	72.1 $\pm$ 0.2	36.4 $\pm$ 0.462.2 $\pm$ 0.2	72.1 $\pm$ 0.2	66.6 $\pm$ 0.468.4 $\pm$ 0.2	78.9 $\pm$ 0.2	67.1 $\pm$ 0.468.4 $\pm$ 0.2	85.9 $\pm$ 0.2
	GCN-BNN	81.0 $\pm$ 0.477.8 $\pm$ 0.2	89.0 $\pm$ 0.2	65.0 $\pm$ 0.462.4 $\pm$ 0.2	84.9 $\pm$ 0.2	63.0 $\pm$ 0.569.2 $\pm$ 0.2	73.1 $\pm$ 0.2	45.9 $\pm$ 0.605.7 $\pm$ 0.2	73.1 $\pm$ 0.2	38.1 $\pm$ 0.580.0 $\pm$ 0.2	73.1 $\pm$ 0.2	69.0 $\pm$ 0.454.4 $\pm$ 0.2	78.9 $\pm$ 0.2	65.1 $\pm$ 0.468.4 $\pm$ 0.2	88.6 $\pm$ 0.2
	GCN-Ems	85.1 $\pm$ 0.467.9 $\pm$ 0.2	90.4 $\pm$ 0.2	73.8 $\pm$ 0.440.1 $\pm$ 0.2	87.3 $\pm$ 0.2	65.6 $\pm$ 0.515.2 $\pm$ 0.2	75.9 $\pm$ 0.2	54.7 $\pm$ 0.553.9 $\pm$ 0.2	75.9 $\pm$ 0.2	37.4 $\pm$ 0.454.7 $\pm$ 0.2	75.9 $\pm$ 0.2	71.8 $\pm$ 0.563.2 $\pm$ 0.2	82.7 $\pm$ 0.2	69.2 $\pm$ 0.465.8 $\pm$ 0.2	91.5 $\pm$ 0.2
	GCN	77.1 $\pm$ 0.482.2 $\pm$ 0.2	88.5 $\pm$ 0.2	71.9 $\pm$ 0.462.8 $\pm$ 0.2	86.4 $\pm$ 0.2	53.0 $\pm$ 0.533.2 $\pm$ 0.2	72.8 $\pm$ 0.2	51.4 $\pm$ 0.551.2 $\pm$ 0.2	72.8 $\pm$ 0.2	52.4 $\pm$ 0.557.1 $\pm$ 0.2	72.8 $\pm$ 0.2	55.3 $\pm$ 0.462.0 $\pm$ 0.2	78.4 $\pm$ 0.2	60.2 $\pm$ 0.457.7 $\pm$ 0.2	87.6 $\pm$ 0.2
	SGCN	84.1 $\pm$ 0.483.2 $\pm$ 0.2	90.2 $\pm$ 0.2	78.3 $\pm$ 0.475.3 $\pm$ 0.2	89.9 $\pm$ 0.2	52.7 $\pm$ 0.561.4 $\pm$ 0.2	76.2 $\pm$ 0.2	42.2 $\pm$ 0.637.3 $\pm$ 0.2	76.2 $\pm$ 0.2	37.1 $\pm$ 0.551.5 $\pm$ 0.2	76.2 $\pm$ 0.2	73.3 $\pm$ 0.457.0 $\pm$ 0.2	74.8 $\pm$ 0.2	66.3 $\pm$ 0.470.6 $\pm$ 0.2	90.0 $\pm$ 0.2
	GCN-EBM	85.0 $\pm$ 0.485.5 $\pm$ 0.2	90.3 $\pm$ 0.2	73.7 $\pm$ 0.475.1 $\pm$ 0.2	87.4 $\pm$ 0.2	66.0 $\pm$ 0.63.8 $\pm$ 0.2	75.8 $\pm$ 0.2	55.1 $\pm$ 0.553.9 $\pm$ 0.2	75.8 $\pm$ 0.2	37.9 $\pm$ 0.455.9 $\pm$ 0.2	75.8 $\pm$ 0.2	71.9 $\pm$ 0.471.6 $\pm$ 0.2	82.6 $\pm$ 0.2	69.1 $\pm$ 0.465.8 $\pm$ 0.2	91.2 $\pm$ 0.2
	GCN-HEAT	85.0 $\pm$ 0.479.5 $\pm$ 0.2	90.3 $\pm$ 0.2	73.7 $\pm$ 0.470.4 $\pm$ 0.2	87.4 $\pm$ 0.2	66.0 $\pm$ 0.561.2 $\pm$ 0.2	75.8 $\pm$ 0.2	55.1 $\pm$ 0.559.1 $\pm$ 0.2	75.8 $\pm$ 0.2	37.9 $\pm$ 0.473.7 $\pm$ 0.2	75.8 $\pm$ 0.2	71.9 $\pm$ 0.467.3 $\pm$ 0.2	82.6 $\pm$ 0.2	66.0 $\pm$ 0.463.8 $\pm$ 0.2	91.2 $\pm$ 0.2
	GCN-GEM	85.0 $\pm$ 0.487.3 $\pm$ 0.2	90.3 $\pm$ 0.2	73.7 $\pm$ 0.476.5 $\pm$ 0.2	87.4 $\pm$ 0.2	66.0 $\pm$ 0.555.2 $\pm$ 0.2	75.8 $\pm$ 0.2	55.1 $\pm$ 0.551.0 $\pm$ 0.2	75.8 $\pm$ 0.2	37.9 $\pm$ 0.439.6 $\pm$ 0.2	75.8 $\pm$ 0.2	71.9 $\pm$ 0.454.0 $\pm$ 0.2	82.6 $\pm$ 0.2	66.0 $\pm$ 0.463.7 $\pm$ 0.2	91.2 $\pm$ 0.2
CoraML	GCN-MCD	84.0 $\pm$ 0.496.4 $\pm$ 0.2	90.1 $\pm$ 0.2	74.0 $\pm$ 0.432.1 $\pm$ 0.2	89.5 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.562.9 $\pm$ 0.2	70.9 $\pm$ 0.2	41.1 $\pm$ 0.477.8 $\pm$ 0.2	70.9 $\pm$ 0.2	70.0 $\pm$ 0.453.3 $\pm$ 0.2	84.0 $\pm$ 0.2	65.4 $\pm$ 0.469.6 $\pm$ 0.2	91.9 $\pm$ 0.2
	GCN-DE	80.7 $\pm$ 0.495.2 $\pm$ 0.2	87.8 $\pm$ 0.2	72.2 $\pm$ 0.445.7 $\pm$ 0.2	87.9 $\pm$ 0.2	n.a/n.a	n.a	75.1 $\pm$ 0.537.9 $\pm$ 0.2	75.0 $\pm$ 0.2	39.9 $\pm$ 0.464.8 $\pm$ 0.2	75.0 $\pm$ 0.2	65.0 $\pm$ 0.467.8 $\pm$ 0.2	81.8 $\pm$ 0.2	64.8 $\pm$ 0.451.3 $\pm$ 0.2	88.5 $\pm$ 0.2
	GCN-MCD+DE	80.3 $\pm$ 0.458.3 $\pm$ 0.2	88.1 $\pm$ 0.2	72.6 $\pm$ 0.446.3 $\pm$ 0.2	87.9 $\pm$ 0.2	n.a/n.a	n.a	77.5 $\pm$ 0.537.8 $\pm$ 0.2	75.2 $\pm$ 0.2	39.7 $\pm$ 0.464.6 $\pm$ 0.2	75.2 $\pm$ 0.2	64.7 $\pm$ 0.469.3 $\pm$ 0.2	81.3 $\pm$ 0.2	64.8 $\pm$ 0.452.7 $\pm$ 0.2	88.7 $\pm$ 0.2
	GCN-BNN	80.3 $\pm$ 0.467.5 $\pm$ 0.2	90.0 $\pm$ 0.2	72.6 $\pm$ 0.461.2 $\pm$ 0.2	89.3 $\pm$ 0.2	n.a/n.a	n.a	72.8 $\pm$ 0.440.8 $\pm$ 0.2	76.8 $\pm$ 0.2	41.7 $\pm$ 0.465.7 $\pm$ 0.2	76.8 $\pm$ 0.2	68.5 $\pm$ 0.452.2 $\pm$ 0.2	83.7 $\pm$ 0.2	66.2 $\pm$ 0.467.7 $\pm$ 0.2	90.9 $\pm$ 0.2
	GCN-Ems	83.9 $\pm$ 0.470.1 $\pm$ 0.2	90.1 $\pm$ 0.2	74.1 $\pm$ 0.457.5 $\pm$ 0.2	89.5 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.446.2 $\pm$ 0.2	77.3 $\pm$ 0.2	39.9 $\pm$ 0.465.6 $\pm$ 0.2	77.3 $\pm$ 0.2	70.2 $\pm$ 0.459.9 $\pm$ 0.2	84.2 $\pm$ 0.2	65.4 $\pm$ 0.462.2 $\pm$ 0.2	91.3 $\pm$ 0.2
	GCN	75.5 $\pm$ 0.483.3 $\pm$ 0.2	89.4 $\pm$ 0.2	75.1 $\pm$ 0.469.5 $\pm$ 0.2	88.1 $\pm$ 0.2	n.a/n.a	n.a	53.9 $\pm$ 0.555.2 $\pm$ 0.2	75.8 $\pm$ 0.2	52.3 $\pm$ 0.580.0 $\pm$ 0.2	75.8 $\pm$ 0.2	55.6 $\pm$ 0.462.5 $\pm$ 0.2	81.3 $\pm$ 0.2	63.0 $\pm$ 0.447.0 $\pm$ 0.2	90.9 $\pm$ 0.2
	SGCN	78.5 $\pm$ 0.481.8 $\pm$ 0.2	89.8 $\pm$ 0.2	72.8 $\pm$ 0.469.6 $\pm$ 0.2	89.8 $\pm$ 0.2	n.a/n.a	n.a	64.4 $\pm$ 0.551.7 $\pm$ 0.2	68.1 $\pm$ 0.2	40.8 $\pm$ 0.444.9 $\pm$ 0.2	68.1 $\pm$ 0.2	68.1 $\pm$ 0.460.7 $\pm$ 0.2	78.9 $\pm$ 0.2	59.8 $\pm$ 0.461.8 $\pm$ 0.2	80.2 $\pm$ 0.2
	GCN-EBM	83.7 $\pm$ 0.484.0 $\pm$ 0.2	90.2 $\pm$ 0.2	74.1 $\pm$ 0.475.9 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.465.1 $\pm$ 0.2	77.1 $\pm$ 0.2	39.9 $\pm$ 0.439.2 $\pm$ 0.2	77.1 $\pm$ 0.2	69.9 $\pm$ 0.469.1 $\pm$ 0.2	83.7 $\pm$ 0.2	65.3 $\pm$ 0.466.5 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-HEAT	83.7 $\pm$ 0.481.1 $\pm$ 0.2	90.2 $\pm$ 0.2	74.1 $\pm$ 0.468.0 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.469.1 $\pm$ 0.2	77.1 $\pm$ 0.2	39.9 $\pm$ 0.467.0 $\pm$ 0.2	77.1 $\pm$ 0.2	69.9 $\pm$ 0.458.8 $\pm$ 0.2	83.7 $\pm$ 0.2	65.3 $\pm$ 0.462.7 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-GEM	83.7 $\pm$ 0.484.2 $\pm$ 0.2	90.2 $\pm$ 0.2	74.1 $\pm$ 0.476.8 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.472.8 $\pm$ 0.2	77.1 $\pm$ 0.2	39.9 $\pm$ 0.487.3 $\pm$ 0.2	77.1 $\pm$ 0.2	69.9 $\pm$ 0.453.2 $\pm$ 0.2	83.7 $\pm$ 0.2	65.3 $\pm$ 0.467.2 $\pm$ 0.2	91.1 $\pm$ 0.2
	GCN-GEM	83.7 $\pm$ 0.480.4 $\pm$ 0.2	90.2 $\pm$ 0.2	74.1 $\pm$ 0.478.5 $\pm$ 0.2	89.4 $\pm$ 0.2	n.a/n.a	n.a	70.3 $\pm$ 0.475.5 $\pm$ 0.2	77.1 $\pm$ 0.2	39.9 $\pm$ 0.489.6 $\pm$ 0.2	77.1 $\pm$ 0.2	69.9 $\pm$ 0.455.5 $\pm$ 0.2	83.7 $\pm$ 0.2	65.3 $\pm$ 0.469.9 $\pm$ 0.2	91.1 $\pm$ 0.2
Citeseer	GCN-MCD	53.6 $\pm$ 0.441.6 $\pm$ 0.2	92.3 $\pm$ 0.2	35.3 $\pm$ 0.471.9 $\pm$ 0.2	82.2 $\pm$ 0.2	n.a/n.a	n.a	54.8 $\pm$ 0.501.4 $\pm$ 0.2	76.9 $\pm$ 0.2	38.5 $\pm$ 0.490.8 $\pm$ 0.2	76.9 $\pm$ 0.2	62.2 $\pm$ 0.460.9 $\pm$ 0.2	78.7 $\pm$ 0.2	54.8 $\pm$ 0.496.9 $\pm$ 0.2	83.3 $\pm$ 0.2
	GCN-DE	53.0 $\pm$ 0.443.5 $\pm$ 0.2	91.8 $\pm$ 0.2	29.8 $\pm$ 0.424.3 $\pm$ 0.2	79.2 $\pm$ 0.2	n.a/n.a	n.a	51.3 $\pm$ 0.493.3 $\pm$ 0.2	74.9 $\pm$ 0.2	36.9 $\pm$ 0.458.2 $\pm$ 0.2	74.9 $\pm$ 0.2	55.4 $\pm$ 0.465.5 $\pm$ 0.2	76.8 $\pm$ 0.2	55.0 $\pm$ 0.472.0 $\pm$ 0.2	89.9 $\pm$ 0.2
	GCN-MCD+DE	52.3 $\pm$ 0.443.5 $\pm$ 0.2	91.7 $\pm$ 0.2	30.5 $\pm$ 0.424.0 $\pm$ 0.2	79.3 $\pm$ 0.2	n.a/n.a	n.a	51.5 $\pm$ 0.493.0 $\pm$ 0.2	74.9 $\pm$ 0.2	36.9 $\pm$ 0.458.0 $\pm$ 0.2	74.9 $\pm$ 0.2	55.4 $\pm$ 0.465.5 $\pm$ 0.2	77.1 $\pm$ 0.2	55.1 $\pm$ 0.472.0 $\pm$ 0.2	89.7 $\pm$ 0.2
	GCN-BNN	52.5 $\pm$ 0.450.3 $\pm$ 0.2	92.7 $\pm$ 0.2	32.8 $\pm$ 0.431.5 $\pm$ 0.2	81.7 $\pm$ 0.2	n.a/n.a	n.a	50.5 $\pm$ 0.490.0 $\pm$ 0.2	76.4 $\pm$ 0.2	37.9 $\pm$ 0.462.0 $\pm$ 0.2	76.4 $\pm$ 0.2	60.3 $\pm$ 0.464.7 $\pm$ 0.2	77.1 $\pm$ 0.2	52.4 $\pm$ 0.459.3 $\pm$ 0.2	82.8 $\pm$ 0.2
	GCN-Ems	53.6 $\pm$ 0.446.7 $\pm$ 0.2	92.9 $\pm$ 0.2	33.5 $\pm$ 0.430.9 $\pm$ 0.2	82.0 $\pm$ 0.2	n.a/n.a	n.a	54.7 $\pm$ 0.496.7 $\pm$ 0.2	76.4 $\pm$ 0.2	37.9 $\pm$ 0.462.0 $\pm$ 0.2	76.4 $\pm$ 0.2	60.3 $\pm$ 0.464.7 $\pm$ 0.2	77.1 $\pm$ 0.2	52.4 $\pm$ 0.459.3 $\pm$ 0.2	82.8 $\pm$ 0.2
	GCN	50.1 $\pm$ 0.453.5 $\pm$ 0.2	92.9 $\pm$ 0.2	34.0 $\pm$ 0.439.4 $\pm$ 0.2	85.0 $\pm$ 0.2	n.a/n.a	n.a	54.7 $\pm$ 0.511.9 $\pm$ 0.2	76.9 $\pm$ 0.2	54.1 $\pm$ 0.477.0 $\pm$ 0.2	76.9 $\pm$ 0.2	77.7 $\pm$ 0.461.3 $\pm$ 0.2	80.2 $\pm$ 0.2	56.9 $\pm$ 0.455.7 $\pm$ 0.2	88.2 $\pm$ 0.2
	SGCN	52.4 $\pm$ 0.456.6 $\pm$ 0.2	92.7 $\pm$ 0.2	32.4 $\pm$ 0.435.4 $\pm$ 0.2	83.8 $\pm$ 0.2	n.a/n.a	n.a	58.4 $\pm$ 0.452.2 $\pm$ 0.2	77.2 $\pm$ 0.2	38.2 $\pm$ 0.439.5 $\pm$ 0.2	77.2 $\pm$ 0.2	64.2 $\pm$ 0.470.2 $\pm$ 0.2	80.8 $\pm$ 0.2	55.1 $\pm$ 0.454.0 $\pm$ 0.2	83.7 $\pm$ 0.2
	GCN-EBM	53.6 $\pm$ 0.450.9 $\pm$ 0.2	92.3 $\pm$ 0.2	32.7 $\pm$ 0.436.7 $\pm$ 0.2	82.2 $\pm$ 0.2	n.a/n.a	n.a	54.7 $\pm$ 0.466.7 $\pm$ 0.2	76.9 $\pm$ 0.2	38.2 $\pm$ 0.439.5 $\pm$ 0.2	76.9 $\pm$ 0.2	62.3 $\pm$ 0.459.1 $\pm$ 0.2	78.6 $\pm$ 0.2	54.8 $\pm$ 0.457.9 $\pm$ 0.2	83.3 $\pm$ 0.2
	GCN-HEAT	53.6 $\pm$ 0.450.9 $\pm$ 0.2	92.3 $\pm$ 0.2	33.7 $\pm$ 0.438.1 $\pm$ 0.2	82.2 $\pm$ 0.2	n.a/n.a	n.a	54.7 $\pm$ 0.450.8 $\pm$ 0.2	76.9 $\pm$ 0.2	38.2 $\pm$ 0.439.5 $\pm$ 0.2	76.9 $\pm$ 0.2	62.3 $\pm$ 0.459.1 $\pm$ 0.2	78.6 $\pm$ 0.2	54.8 $\pm$ 0.457.9 $\pm$ 0.2	83.3 $\pm$ 0.2
	GCN-GEM	53.6 $\pm$ 0.456.7 $\pm$ 0.2	92.3 $\pm$ 0.2	33.7 $\pm$ 0.440.5 $\pm$ 0.2	82.2 $\pm$ 0.2	n.a/n.a	n.a	54.7 $\pm$ 0.484.9 $\pm$ 0.2	76.9 $\pm$ 0.2	38.2 $\pm$ 0.442.5 $\pm$ 0.2	76.9 $\pm$ 0.2	62.3 $\pm$ 0.459.1 $\pm$ 0.2	78.6 $\pm$ 0.2	54.8 $\pm$ 0.457.9 $\pm$ 0.2	83.3 $\pm$ 0.2
	GCN-GEM	53.6 $\pm$ 0.456.7 $\pm$ 0.2	92.3 $\pm$ 0.2	33.7 $\pm$ 0.440.5 $\pm$ 0.2	82.2										

Model	CoraML	CoraML LLM	Citeseer	PubMed	Amazon Computers	Amazon Photo	Coauthor CS	Coauthor Physics
GCN-MCD	9.7/18.9	9.2/19.0	9.4/18.7	9.1/18.7	9.1/18.3	9.9/19.1	10.0/19.2	10.5/20.4
GCN-DE	6.3/14.1	7.9/16.9	7.2/15.3	7.1/13.3	7.4/16.4	6.9/15.9	6.3/13.2	7.8/15.4
GCN-MCD+DE	6.8/14.7	7.2/15.9	6.0/13.9	6.6/12.8	7.9/16.9	7.6/16.6	6.7/13.5	7.4/15.1
GCN-BNN	5.9/12.3	6.6/14.1	5.6/11.8	5.8/11.4	4.6/10.4	5.3/9.7	5.2/11.6	5.3/11.9
GCN-Ens	7.4/15.7	6.8/16.0	6.8/13.5	7.2/13.4	6.4/14.9	4.4/8.9	4.9/11.1	6.4/13.6
GPN	7.1/15.3	6.0/14.8	6.4/13.1	4.8/9.0	5.4/12.9	5.3/11.8	6.8/13.5	5.6/12.5
SGCN	4.7/9.1	5.9/14.8	5.8/13.1	4.2/7.4	9.4/18.7	7.4/14.8	9.4/18.6	4.6/9.9
GCN-EBM	4.2/8.6	3.8/7.6	4.8/11.9	6.5/12.5	3.4/8.7	4.6/10.7	3.8/6.6	4.2/8.7
GCN-HEAT	4.8/11.8	4.6/13.1	4.7/11.0	5.5/12.7	3.6/10.9	4.7/12.4	4.6/10.6	5.2/13.7
GCNSafe	6.4/11.9	4.8/10.1	5.6/11.2	5.4/8.9	5.8/13.3	6.9/14.7	5.4/10.3	6.1/11.9
GCN-GEBM	2.7 / 4.5	3.2 / 5.7	3.6 / 7.7	3.6 / 7.0	2.8 / 6.5	3.0 / 5.0	2.8 / 4.7	3.0 / 4.8

Table 11: Average o.o.d. detection rank (AUC-PR) ($\downarrow$) of epistemic uncertainty versus other epistemic measures / all uncertainty measures over all distribution shifts in an inductive setting (**best** and **runner-up**).

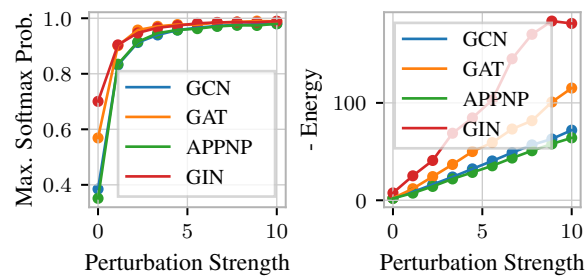


Figure 5: Confidence (Maximum Softmax Probability and negative energy) for different GNNs at increasing distribution shift severity.

of GNNs on in-distribution data, and provides strong evidence for models to be underconfident. This does not conflict with our claim that with increasing distance from the training distribution, i.e. on out-of-distribution data, piece affine GNNs become overconfident. We empirically verify this in Figure 5: Both the energy and the maximum softmax probability - confidence measures - of the GNN architectures visualized increase (or converge to their maximum) under more severe normal feature perturbations.

C.9 Robust Evidential Inference

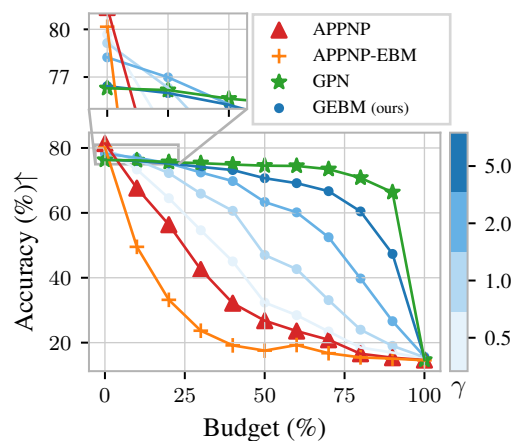


Figure 6: Robust evidential inference using APPNP as a backbone for GEBM at increasingly severe feature perturbations.

	Model	LoC (<i>loot</i>)			LoC (<i>hetero</i>)			Ber(<i>p</i>) (<i>near</i>)			Ber(0.5)			AUC/R1 (<i>far</i>)			Page Rank			Homophily		
		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$		AUC-PR $\uparrow$ (Area / Ep.)	Acc. $\uparrow$	
CoraML	GCN-MCD	82.8 $\pm$ 4.486.1 $\pm$ 0.2	90.4 $\pm$ 0.8		69.5 $\pm$ 3.733.4 $\pm$ 0.3	85.2 $\pm$ 0.4		65.3 $\pm$ 4.051.9 $\pm$ 0.9	82.0 $\pm$ 0.8		56.5 $\pm$ 4.050.3 $\pm$ 0.9	82.0 $\pm$ 0.8		37.4 $\pm$ 1.488.0 $\pm$ 0.7	82.0 $\pm$ 0.8		70.0 $\pm$ 1.056.5 $\pm$ 0.1	82.0 $\pm$ 0.8		63.7 $\pm$ 1.483.3 $\pm$ 0.1	83.0 $\pm$ 0.8	
	GCN-DE	79.4 $\pm$ 3.461.2 $\pm$ 0.3	87.3 $\pm$ 0.3		64.2 $\pm$ 4.041.5 $\pm$ 0.3	80.9 $\pm$ 0.3		60.3 $\pm$ 4.055.9 $\pm$ 0.9	78.4 $\pm$ 1.0		48.3 $\pm$ 4.071.0 $\pm$ 0.3	78.4 $\pm$ 1.0		35.6 $\pm$ 1.059.3 $\pm$ 0.3	78.4 $\pm$ 1.0		64.1 $\pm$ 0.98.1 $\pm$ 0.1	79.0 $\pm$ 0.3		63.4 $\pm$ 0.98.1 $\pm$ 0.1	90.0 $\pm$ 0.8	
	GCN-MCD+DE	79.4 $\pm$ 3.462.2 $\pm$ 0.3	87.6 $\pm$ 0.3		63.3 $\pm$ 4.042.3 $\pm$ 0.3	80.1 $\pm$ 0.3		63.9 $\pm$ 4.055.8 $\pm$ 0.7	78.1 $\pm$ 0.2		48.7 $\pm$ 4.071.2 $\pm$ 0.3	78.1 $\pm$ 0.2		35.8 $\pm$ 1.061.1 $\pm$ 0.3	78.1 $\pm$ 0.2		64.5 $\pm$ 0.97.5 $\pm$ 0.1	79.5 $\pm$ 0.7		63.5 $\pm$ 0.97.5 $\pm$ 0.1	90.0 $\pm$ 0.8	
	GCN-BNN	77.2 $\pm$ 3.477.4 $\pm$ 0.3	80.0 $\pm$ 0.3		58.2 $\pm$ 3.054.3 $\pm$ 0.3	82.1 $\pm$ 0.3		64.1 $\pm$ 0.98.4 $\pm$ 0.3	80.6 $\pm$ 0.3		47.9 $\pm$ 0.965.8 $\pm$ 0.3	80.6 $\pm$ 0.3		37.5 $\pm$ 1.055.2 $\pm$ 0.3	80.6 $\pm$ 0.3		67.6 $\pm$ 0.953.8 $\pm$ 0.3	80.6 $\pm$ 0.3		63.3 $\pm$ 0.964.3 $\pm$ 0.3	93.1 $\pm$ 0.8	
	GCN-Ems	82.8 $\pm$ 4.069.0 $\pm$ 0.3	90.3 $\pm$ 0.3		69.5 $\pm$ 3.042.4 $\pm$ 0.3	85.2 $\pm$ 0.3		66.3 $\pm$ 4.051.3 $\pm$ 0.3	81.8 $\pm$ 0.3		56.0 $\pm$ 4.062.3 $\pm$ 0.3	81.8 $\pm$ 0.3		37.5 $\pm$ 1.054.9 $\pm$ 0.3	81.8 $\pm$ 0.3		70.1 $\pm$ 0.953.3 $\pm$ 0.3	82.1 $\pm$ 0.3		63.6 $\pm$ 0.964.0 $\pm$ 0.3	94.0 $\pm$ 0.8	
	GPIN	76.0 $\pm$ 3.082.3 $\pm$ 0.3	90.0 $\pm$ 0.3		65.0 $\pm$ 4.063.2 $\pm$ 0.3	85.5 $\pm$ 0.3		50.8 $\pm$ 4.053.1 $\pm$ 0.3	78.8 $\pm$ 0.3		49.8 $\pm$ 4.051.4 $\pm$ 0.3	78.8 $\pm$ 0.3		50.7 $\pm$ 0.956.0 $\pm$ 0.3	78.8 $\pm$ 0.3		55.6 $\pm$ 0.950.5 $\pm$ 0.3	78.4 $\pm$ 0.3		67.9$\pm$0.949.1$\pm$0.3	91.0 $\pm$ 0.8	
	SGCN	83.1 $\pm$ 3.086.1 $\pm$ 0.3	91.0 $\pm$ 0.3		73.2$\pm$3.078.9$\pm$0.3	85.5 $\pm$ 0.3		54.9 $\pm$ 0.954.5 $\pm$ 0.3	51.7 $\pm$ 0.3		45.3 $\pm$ 0.962.0 $\pm$ 0.3	51.7 $\pm$ 0.3		36.2 $\pm$ 1.044.3 $\pm$ 0.3	51.7 $\pm$ 0.3		69.3 $\pm$ 0.962.6 $\pm$ 0.3	83.3 $\pm$ 0.3		64.1 $\pm$ 0.965.8 $\pm$ 0.3	94.2 $\pm$ 0.8	
	GCN-EBM	82.8 $\pm$ 4.083.0 $\pm$ 0.3	90.4 $\pm$ 0.8		69.1 $\pm$ 3.069.6 $\pm$ 0.3	85.0 $\pm$ 0.3		66.4 $\pm$ 4.056.1 $\pm$ 0.3	81.9 $\pm$ 0.3		57.1 $\pm$ 0.956.3 $\pm$ 0.3	81.9 $\pm$ 0.3		37.5 $\pm$ 1.055.0 $\pm$ 0.3	81.9 $\pm$ 0.3		70.1 $\pm$ 0.953.0 $\pm$ 0.3	82.5 $\pm$ 0.3		63.4 $\pm$ 0.962.3 $\pm$ 0.3	94.1 $\pm$ 0.8	
	GCN-HEAT	82.8 $\pm$ 4.079.4 $\pm$ 0.3	90.4 $\pm$ 0.8		69.1 $\pm$ 3.063.0 $\pm$ 0.3	85.0 $\pm$ 0.3		66.4$\pm$4.052.9$\pm$0.3	81.9 $\pm$ 0.3		57.1 $\pm$ 0.959.5 $\pm$ 0.3	81.9 $\pm$ 0.3		37.5 $\pm$ 1.073.6 $\pm$ 0.3	81.9 $\pm$ 0.3		70.0 $\pm$ 0.967.9 $\pm$ 0.3	82.5 $\pm$ 0.3		63.4 $\pm$ 0.963.8 $\pm$ 0.3	94.1 $\pm$ 0.8	
	GCNsafe	82.8 $\pm$ 4.085.3 $\pm$ 0.7	90.4 $\pm$ 0.8		69.1 $\pm$ 3.073.2 $\pm$ 0.3	85.0 $\pm$ 0.3		66.4$\pm$4.057.7$\pm$0.3	81.9 $\pm$ 0.3		57.1 $\pm$ 0.952.9 $\pm$ 0.3	81.9 $\pm$ 0.3		37.5 $\pm$ 1.078.5 $\pm$ 0.3	81.9 $\pm$ 0.3		70.0 $\pm$ 0.963.0 $\pm$ 0.3	82.5 $\pm$ 0.3		63.4 $\pm$ 0.960.9 $\pm$ 0.3	94.1 $\pm$ 0.8	
GCN-GERM	82.8$\pm$4.086.1$\pm$0.6	90.4$\pm$0.8		69.1$\pm$3.073.2$\pm$0.3	85.0$\pm$0.3		66.4$\pm$4.071.2$\pm$0.3	81.9$\pm$0.3		57.1$\pm$0.951.6$\pm$0.3	81.9$\pm$0.3		37.5$\pm$1.091.7$\pm$0.3	81.9$\pm$0.3		70.0$\pm$0.963.3$\pm$0.7	82.5$\pm$0.3		63.4$\pm$0.964.3$\pm$0.3	94.1$\pm$0.8		
CoraML-LM	GCN-MCD	84.4 $\pm$ 3.045.7 $\pm$ 0.2	91.2 $\pm$ 0.7		70.7 $\pm$ 3.033.2 $\pm$ 0.3	90.6 $\pm$ 0.3		n.a./n.a.	n.a.		70.2 $\pm$ 3.053.1 $\pm$ 0.3	82.6 $\pm$ 0.3		42.9 $\pm$ 3.049.4 $\pm$ 0.3	82.6 $\pm$ 0.3		70.7$\pm$3.053.6$\pm$0.1	84.9 $\pm$ 0.3		64.8 $\pm$ 3.047.8 $\pm$ 0.3	94.1 $\pm$ 0.8	
	GCN-DE	82.8 $\pm$ 4.061.5 $\pm$ 0.3	89.7 $\pm$ 0.3		69.1 $\pm$ 3.049.2 $\pm$ 0.3	88.9 $\pm$ 0.7		n.a./n.a.	n.a.		75.8$\pm$4.063.1$\pm$0.3	79.9 $\pm$ 0.3		42.2 $\pm$ 3.065.0 $\pm$ 0.3	79.9 $\pm$ 0.3		64.7 $\pm$ 0.989.1 $\pm$ 0.3	82.2 $\pm$ 0.7		64.0 $\pm$ 0.953.3 $\pm$ 0.3	91.7 $\pm$ 0.8	
	GCN-MCD+DE	83.2 $\pm$ 3.061.3 $\pm$ 0.3	89.4 $\pm$ 0.3		68.7 $\pm$ 4.048.8 $\pm$ 0.3	88.9 $\pm$ 0.7		n.a./n.a.	n.a.		75.9$\pm$4.040.7$\pm$0.3	80.3 $\pm$ 0.3		41.1 $\pm$ 0.964.2 $\pm$ 0.3	80.3 $\pm$ 0.3		64.6 $\pm$ 0.969.4 $\pm$ 0.3	82.7 $\pm$ 0.3		63.6 $\pm$ 0.953.8 $\pm$ 0.3	91.4 $\pm$ 0.8	
	GCN-BNN	83.8 $\pm$ 3.072.5 $\pm$ 0.3	91.4 $\pm$ 0.8		69.9 $\pm$ 4.061.9 $\pm$ 0.3	91.0 $\pm$ 0.3		n.a./n.a.	n.a.		72.3 $\pm$ 4.043.2 $\pm$ 0.3	82.8 $\pm$ 0.3		41.9 $\pm$ 1.064.0 $\pm$ 0.3	82.8 $\pm$ 0.3		68.7 $\pm$ 0.951.1 $\pm$ 0.3	84.0 $\pm$ 0.3		67.2$\pm$0.968.8$\pm$0.3	94.0 $\pm$ 0.8	
	GCN-Ems	83.8 $\pm$ 3.077.1 $\pm$ 0.7	91.4 $\pm$ 0.8		70.9 $\pm$ 3.056.0 $\pm$ 0.3	91.0 $\pm$ 0.3		n.a./n.a.	n.a.		70.8 $\pm$ 4.052.3 $\pm$ 0.3	82.6 $\pm$ 0.3		41.1 $\pm$ 0.963.7 $\pm$ 0.3	82.6 $\pm$ 0.3		70.8 $\pm$ 0.978.3 $\pm$ 0.3	85.0 $\pm$ 0.3		64.9 $\pm$ 0.963.4 $\pm$ 0.3	94.0 $\pm$ 0.8	
	GPIN	76.0 $\pm$ 3.085.4 $\pm$ 0.3	90.5 $\pm$ 0.3		73.4 $\pm$ 3.070.9 $\pm$ 0.3	88.5 $\pm$ 0.3		n.a./n.a.	n.a.		53.2 $\pm$ 4.054.1 $\pm$ 0.3	79.7 $\pm$ 0.3		52.3 $\pm$ 0.955.5 $\pm$ 0.3	79.7 $\pm$ 0.3		55.2 $\pm$ 0.961.0 $\pm$ 0.3	80.9 $\pm$ 0.3		71.0$\pm$0.951.3$\pm$0.3	92.5 $\pm$ 0.8	
	SGCN	86.2 $\pm$ 3.088.7 $\pm$ 0.3	90.3 $\pm$ 0.8		74.3 $\pm$ 3.079.3 $\pm$ 0.3	90.6 $\pm$ 0.3		n.a./n.a.	n.a.		64.4 $\pm$ 4.053.0 $\pm$ 0.3	68.7 $\pm$ 0.3		41.1 $\pm$ 0.947.2 $\pm$ 0.3	68.7 $\pm$ 0.3		68.5 $\pm$ 0.964.9 $\pm$ 0.3	82.0 $\pm$ 0.7		63.4 $\pm$ 0.964.4 $\pm$ 0.3	93.1 $\pm$ 0.8	
	GCN-EBM	84.3 $\pm$ 3.084.0 $\pm$ 0.7	91.4 $\pm$ 0.8		70.9 $\pm$ 3.073.0 $\pm$ 0.3	90.7 $\pm$ 0.3		n.a./n.a.	n.a.		70.6 $\pm$ 0.966.0 $\pm$ 0.3	82.6 $\pm$ 0.3		42.3 $\pm$ 1.041.0 $\pm$ 0.3	82.6 $\pm$ 0.3		70.6 $\pm$ 0.966.2 $\pm$ 0.3	85.0 $\pm$ 0.3		65.0 $\pm$ 0.964.7 $\pm$ 0.3	93.8 $\pm$ 0.8	
	GCN-HEAT	84.3 $\pm$ 3.078.0 $\pm$ 0.7	91.4 $\pm$ 0.8		70.9 $\pm$ 3.065.2 $\pm$ 0.3	90.7 $\pm$ 0.3		n.a./n.a.	n.a.		70.6 $\pm$ 0.967.0 $\pm$ 0.3	82.6 $\pm$ 0.3		42.3 $\pm$ 1.066.8 $\pm$ 0.3	82.6 $\pm$ 0.3		70.6 $\pm$ 0.969.0 $\pm$ 0.3	85.0 $\pm$ 0.3		65.0 $\pm$ 0.964.8 $\pm$ 0.3	93.8 $\pm$ 0.8	
	GCNsafe	84.3 $\pm$ 3.086.7 $\pm$ 0.8	91.4 $\pm$ 0.8		70.9 $\pm$ 3.079.8 $\pm$ 0.3	90.7 $\pm$ 0.3		n.a./n.a.	n.a.		70.6 $\pm$ 0.957.4 $\pm$ 0.3	82.6 $\pm$ 0.3		42.3 $\pm$ 1.045.2 $\pm$ 0.3	82.6 $\pm$ 0.3		70.6 $\pm$ 0.954.9 $\pm$ 0.3	85.0 $\pm$ 0.3		65.0 $\pm$ 0.961.9 $\pm$ 0.3	93.8 $\pm$ 0.8	
GCN-GERM	84.3$\pm$3.085.9$\pm$0.8	91.4$\pm$0.8		70.9$\pm$3.074.6$\pm$0.3	90.7$\pm$0.3		70.9$\pm$3.074.6$\pm$0.3	90.7$\pm$0.3		70.6$\pm$0.972.4$\pm$0.3	82.6$\pm$0.3		42.3$\pm$1.088.4$\pm$0.3	82.6$\pm$0.3		70.6$\pm$0.952.5$\pm$0.3	85.0$\pm$0.3		65.0$\pm$0.965.2$\pm$0.3	93.8$\pm$0.8		
Citeseer	GCN-MCD	66.9 $\pm$ 3.090.4 $\pm$ 0.3	83.5 $\pm$ 0.3		73.9 $\pm$ 4.037.0 $\pm$ 0.3	91.1 $\pm$ 0.8		62.5 $\pm$ 3.051.2 $\pm$ 0.3	82.1 $\pm$ 0.8		50.2 $\pm$ 3.051.3 $\pm$ 0.3	82.1 $\pm$ 0.8		38.4 $\pm$ 3.049.0 $\pm$ 0.7	82.1 $\pm$ 0.8		60.3$\pm$3.050.7$\pm$0.3	82.9 $\pm$ 0.3		51.5 $\pm$ 0.950.0 $\pm$ 0.3	86.3 $\pm$ 0.8	
	GCN-DE	62.6 $\pm$ 3.042.3 $\pm$ 0.3	80.1 $\pm$ 0.7		68.4 $\pm$ 3.051.4 $\pm$ 0.3	86.4 $\pm$ 0.3		61.1 $\pm$ 3.059.2 $\pm$ 0.3	80.7 $\pm$ 0.3		46.3 $\pm$ 3.080.3 $\pm$ 0.3	76.9 $\pm$ 0.3		35.8 $\pm$ 3.064.9 $\pm$ 0.3	76.9 $\pm$ 0.3		56.2 $\pm$ 0.955.9 $\pm$ 0.3	79.0 $\pm$ 0.3		50.6 $\pm$ 0.950.3 $\pm$ 0.3	81.8 $\pm$ 0.8	
	GCN-MCD+DE	62.6 $\pm$ 3.042.3 $\pm$ 0.3	80.1 $\pm$ 0.7		68.4 $\pm$ 3.052.4 $\pm$ 0.3	86.2 $\pm$ 0.3		61.2 $\pm$ 3.056.7 $\pm$ 0.3	78.8 $\pm$ 0.3		47.3 $\pm$ 3.081.4 $\pm$ 0.3	76.8 $\pm$ 0.3		35.8 $\pm$ 3.064.6 $\pm$ 0.3	76.8 $\pm$ 0.3		56.2 $\pm$ 0.956.0 $\pm$ 0.3	79.2 $\pm$ 0.3		50.7 $\pm$ 0.950.3 $\pm$ 0.3	81.7 $\pm$ 0.8	
	GCN-BNN	63.0 $\pm$ 3.058.3 $\pm$ 0.3	82.5 $\pm$ 0.3		70.0 $\pm$ 3.066.6 $\pm$ 0.3	90.2 $\pm$ 0.7		64.7$\pm$3.057.1$\pm$0.3	81.2 $\pm$ 0.3		50.2 $\pm$ 3.072.4 $\pm$ 0.3	81.2 $\pm$ 0.3		37.2 $\pm$ 0.959.9 $\pm$ 0.3	81.2 $\pm$ 0.3		58.6 $\pm$ 0.948.6 $\pm$ 0.3	80.7 $\pm$ 0.3		51.9 $\pm$ 0.953.1 $\pm$ 0.3	86.6 $\pm$ 0.8	
	GCN-Ems	67.6 $\pm$ 3.051.1 $\pm$ 0.3	83.8 $\pm$ 0.3		74.1 $\pm$ 4.064.2 $\pm$ 0.3	91.2 $\pm$ 0.8		63.2 $\pm$ 3.058.1 $\pm$ 0.3	82.0 $\pm$ 0.3		57.0 $\pm$ 3.079.8 $\pm$ 0.3	82.0 $\pm$ 0.3		38.9 $\pm$ 3.058.1 $\pm$ 0.3	82.0 $\pm$ 0.3		60.4$\pm$4.055.3$\pm$0.3	82.6 $\pm$ 0.3		51.5 $\pm$ 0.954.9 $\pm$ 0.3	86.4 $\pm$ 0.8	
	GPIN	66.9 $\pm$ 3.090.4 $\pm$ 0.3	83.5 $\pm$ 0.3		73.9 $\pm$ 4.037.0 $\pm$ 0.3	91.1 $\pm$ 0.8		57.1 $\pm$ 3.059.0 $\pm$ 0.3	83.1 $\pm$ 0.3		52.3 $\pm$ 3.057.0 $\pm$ 0.3	83.1 $\pm$ 0.3	</									

Model	CoraML	CoraML LLM	Citeseer	PubMed	Amazon Computers	Amazon Photo	Coauthor CS	Coauthor Physics
GCN-MCD	9.9/18.9	9.4/19.0	9.1/18.5	9.5/19.1	8.9/18.2	10.1/19.4	10.0/18.9	10.5/20.3
GCN-DE	6.3/13.9	7.4/16.1	7.2/15.6	6.8/12.9	7.6/16.7	7.3/16.4	6.7/13.3	7.1/14.4
GCN-MCD+DE	6.7/14.5	7.5/16.2	6.9/15.3	6.1/12.3	7.6/16.7	6.2/15.3	6.0/12.7	7.2/14.7
GCN-BNN	5.7/12.1	6.7/14.4	6.2/12.8	5.6/12.8	5.3/12.6	4.9/8.9	6.0/12.6	5.7/12.9
GCN-Ens	7.2/14.2	6.8/16.1	6.3/12.9	7.6/13.9	7.2/16.1	5.9/12.3	4.5/11.0	6.4/13.8
GPN	7.0/15.1	6.2/12.9	7.0/15.4	5.1/8.4	5.1/11.2	5.4/11.6	6.6/12.0	5.9/9.8
SGCN	3.9/6.9	4.2/8.8	4.2/8.4	3.5/6.3	7.6/16.5	6.3/13.7	9.3/18.4	4.7/8.5
GCN-EBM	4.8/9.3	4.4/11.2	4.8/13.1	6.4/13.3	3.8/11.3	4.6/11.3	4.0/7.1	4.1/8.6
GCN-HEAT	4.7/10.2	4.4/12.9	3.5/9.8	5.2/12.7	3.9/11.3	4.9/14.1	4.4/10.7	5.2/14.0
GCNSafe	6.7/12.1	5.3/10.8	6.7/12.5	5.8/9.5	6.4/14.7	6.5/15.6	5.7/10.9	6.1/11.8
GCN-GEBM	3.2 / 5.2	3.6 / 6.5	4.2 / 9.3	4.4 / 8.5	2.7 / 6.7	3.9 / 8.6	2.8 / 5.8	3.2 / 5.2

Table 13: Average o.o.d. detection rank (AUC-PR) ($\downarrow$) of epistemic uncertainty versus other epistemic measures / all uncertainty measures over all distribution shifts in a transductive setting (**best** and **runner-up**).

Model	GCN MCD	GCN DE	GCN MCD+DE	GCN BNN	GCN Ens	GPN	SGCN	GCN EBM	GCN HEAT	GCNSafe	GCN GEBM
Global Rank $\downarrow$	19.6	15.3	15.2	10.7	12.6	13.0	13.9	9.4	10.3	10.6	5.0

Table 14: Global o.o.d. detection rank of each model, averaged over all datasets and distribution shifts in an inductive setting based on AUC-ROC scores (**best** and **runner-up**).

We compare the AUC-ROC of different uncertainty estimators while increasing the fraction of perturbed features for each o.o.d. node in Figure 7. While both the softmax-level uncertainty of the vanilla GCN and the EBM again become overconfident, GEBM can reliably identify this distribution shift. This coincides with observations made for feature perturbations that affect all node features simultaneously and justifies focusing on that distribution shift for the bulk of our study.

D Ablations

D.1 Energy at Different Scales

We report AUC-ROC and AUC-PR for o.o.d.-detection ablating GEBM in an inductive and transductive setting with standard deviations in Tables 26 to 29. Additionally, we rank all variants of GEBM (i.e. energies at different structural scales) similar to the o.o.d. detection experiments in Section 5.2. We observe our proposed GEBM to consistently be the most effective over different shifts simultaneously. We can also confirm the effectiveness of the scale-specific energies at detecting distribution shifts that should be covered from their definition. Both the best and second-best ranking methods are GEBM and a variant that uses unregularized energy: This shows the efficacy of a scale-aware EBM for graph problems.

D.2 Diffusion Process

We also ablate the diffusion operator P_A and its hyperparameters, i.e. the type of diffusion process (see Appendix C.6), the number of diffusion steps t as well the teleport probability α , in Figure 8. We evaluate the performance of GEBM on a representative of each distribution shift family. As expected, low t and high α aid node-level anomaly detection while the opposite holds for cluster and local shifts. Label-Propagation achieves satisfactory performance over the entire range of diffusion hyperparameters which justifies using it in our experiments. In particular, we did not tune any hyperparameters for good o.o.d.-detection. As stated in Appendix C.6, it does not bias the energy toward high-degree nodes which explains its advantages over the other diffusion types. It performs well over a broad range of hyperparameters making GEBM less sensitive to those.

Dataset	LoC (<i>last</i>) Abs. / Rel.	LoC (<i>hetero</i>) Abs. / Rel.	Ber($\hat{p}$) (<i>near</i>) Abs. / Rel.	Ber(0.5) Abs. / Rel.	$N(0,1)$ (<i>far</i>) Abs. / Rel.	Page Rank Abs. / Rel.	Homophily Abs. / Rel.	(Weighted) Avg. Abs. / Rel.
CoraML	+1.7 / +1.9%	+1.0 / +1.2%	+10.4 / +15.5%	+35.9 / +61.3%	+60.6 / +235.7%	-16.5 / -22.3%	+5.5 / +7.7%	+10.5 / +32.8%
CoraML-LLM	+1.3 / +1.5%	+2.2 / +2.5%	n.a.	+12.2 / +17.5%	+52.3 / +158.2%	-20.8 / -28.6%	+5.3 / +7.6%	+5.8 / +18.8%
Citeseer	+1.5 / +1.7%	+0.8 / +0.9%	+1.2 / +1.9%	+52.5 / +119.9%	+48.6 / +207.1%	-18.1 / -27.3%	+1.7 / +3.0%	+9.0 / +33.0%
PubMed	+5.1 / +7.8%	-0.6 / -0.9%	n.a.	+35.4 / +72.7%	+48.2 / +175.3%	-23.2 / -31.9%	+6.7 / +12.9%	+8.2 / +28.7%
Amazon Computers	+5.9 / +7.8%	+2.4 / +2.8%	+28.7 / +50.7%	+20.5 / +37.3%	+54.0 / +134.9%	-32.7 / -38.3%	+2.9 / +4.7%	+7.9 / +20.9%
Amazon Photo	+8.4 / +11.2%	+9.4 / +12.0%	+26.9 / +46.6%	+12.2 / +23.4%	+49.8 / +116.8%	-32.3 / -35.7%	+8.4 / +14.0%	+8.9 / +21.0%
Coauthor-CS	+2.9 / +3.2%	+1.9 / +2.0%	+26.0 / +37.3%	+28.0 / +39.1%	+61.2 / +212.8%	-25.4 / -31.0%	+9.0 / +14.8%	+10.9 / +30.3%
Coauthor-Physics	+1.6 / +1.7%	+1.5 / +1.5%	+22.4 / +37.7%	+32.9 / +52.6%	+33.2 / +84.2%	-21.9 / -26.7%	+4.8 / +8.6%	+7.5 / +16.9%

Table 15: Improvement of GEBM in o.o.d. detection AUC-ROC over a vanilla logit-based EBM (GCN-EBM) in an inductive setting.

	Model	LoC (<i>last</i>)		LoC (<i>hetero</i>)		Ber($\hat{p}$) (<i>near</i>)		Ber(0.5)		N(0,1) (<i>far</i>)		Page Rank		Homophily	
		AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$	AUC-ROC $\uparrow$ (Alea. / Epi.)	Acc. $\uparrow$
CoraML	GCN-MCD	91.6 $\pm$ 0.5/0.0 $\pm$ 0.3	90.4 $\pm$ 0.0	89.2 $\pm$ 0.4/49.0 $\pm$ 0.5	87.4 $\pm$ 0.3	77.2 $\pm$ 0.5/50.4 $\pm$ 0.7	75.6 $\pm$ 0.3	75.6 $\pm$ 0.3	45.3 $\pm$ 0.5/50.3 $\pm$ 0.7	75.6 $\pm$ 0.3	80.4 $\pm$ 0.5/50.0 $\pm$ 0.3	82.3 $\pm$ 0.1	85.6 $\pm$ 0.5/50.7 $\pm$ 0.5	91.4 $\pm$ 0.1	85.6 $\pm$ 0.5/50.7 $\pm$ 0.5
	GCN-DE	88.4 $\pm$ 0.7/6.0 $\pm$ 0.3	87.2 $\pm$ 0.7	85.3 $\pm$ 0.7/72.3 $\pm$ 0.3	82.0 $\pm$ 0.3	67.8 $\pm$ 0.7/67.8 $\pm$ 0.3	71.9 $\pm$ 0.7	64.7 $\pm$ 0.7/70.8 $\pm$ 0.3	71.9 $\pm$ 0.7	40.5 $\pm$ 0.5/58.4 $\pm$ 0.3	71.9 $\pm$ 0.7	77.7 $\pm$ 0.7/71.6 $\pm$ 0.3	79.3 $\pm$ 0.1	80.7 $\pm$ 0.6/68.4 $\pm$ 0.3	
	GCN-MCD+DE	88.5 $\pm$ 0.7/6.0 $\pm$ 0.3	87.3 $\pm$ 0.7	85.4 $\pm$ 0.7/72.3 $\pm$ 0.3	83.0 $\pm$ 0.3	67.7 $\pm$ 0.7/69.2 $\pm$ 0.3	72.1 $\pm$ 0.7	63.9 $\pm$ 0.7/69.6 $\pm$ 0.3	72.1 $\pm$ 0.7	41.3 $\pm$ 0.5/59.1 $\pm$ 0.3	72.1 $\pm$ 0.7	77.8 $\pm$ 0.7/70.9 $\pm$ 0.3	79.9 $\pm$ 0.1	80.8 $\pm$ 0.6/75.2 $\pm$ 0.3	
	GCN-BNN	90.1 $\pm$ 0.8/9.1 $\pm$ 0.3	89.0 $\pm$ 0.1	87.3 $\pm$ 0.8/86.5 $\pm$ 0.3	84.9 $\pm$ 0.3	75.7 $\pm$ 0.7/71.7 $\pm$ 0.3	73.1 $\pm$ 0.7	75.7 $\pm$ 0.7/74.8 $\pm$ 0.3	73.1 $\pm$ 0.7	48.7 $\pm$ 0.5/55.0 $\pm$ 0.3	73.1 $\pm$ 0.7	81.3 $\pm$ 0.8/80.7 $\pm$ 0.3	78.9 $\pm$ 0.1	87.0 $\pm$ 0.8/85.0 $\pm$ 0.3	
	GCN-Ens	90.0 $\pm$ 0.7/9.1 $\pm$ 0.3	90.4 $\pm$ 0.0	89.1 $\pm$ 0.6/6.5 $\pm$ 0.3	87.3 $\pm$ 0.3	74.9 $\pm$ 0.8/48.1 $\pm$ 0.3	75.9 $\pm$ 0.3	76.1 $\pm$ 0.6/67.1 $\pm$ 0.3	75.9 $\pm$ 0.3	45.2 $\pm$ 0.5/54.9 $\pm$ 0.3	75.9 $\pm$ 0.3	80.2 $\pm$ 0.6/64.1 $\pm$ 0.3	82.7 $\pm$ 0.1	86.2 $\pm$ 0.6/69.9 $\pm$ 0.3	
	CPN	89.6 $\pm$ 0.8/6.0 $\pm$ 0.3	88.5 $\pm$ 0.1	87.1 $\pm$ 0.8/82.4 $\pm$ 0.3	86.4 $\pm$ 0.7	81.8 $\pm$ 0.7/67.7 $\pm$ 0.3	72.8 $\pm$ 0.3	82.5 $\pm$ 0.6/58.5 $\pm$ 0.3	72.8 $\pm$ 0.3	42.7 $\pm$ 0.5/67.3 $\pm$ 0.3	72.8 $\pm$ 0.3	81.9 $\pm$ 0.8/67.5 $\pm$ 0.3	78.4 $\pm$ 0.1	87.6 $\pm$ 0.8/85.6 $\pm$ 0.3	
	SGCN	91.2 $\pm$ 0.8/9.3 $\pm$ 0.3	90.2 $\pm$ 0.0	90.9 $\pm$ 0.8/88.3 $\pm$ 0.3	89.0 $\pm$ 0.1	69.2 $\pm$ 0.5/51.2 $\pm$ 0.3	36.2 $\pm$ 0.1	65.4 $\pm$ 0.5/44.1 $\pm$ 0.3	36.2 $\pm$ 0.1	49.2 $\pm$ 0.4/49.6 $\pm$ 0.3	36.2 $\pm$ 0.1	79.4 $\pm$ 0.6/69.4 $\pm$ 0.3	74.8 $\pm$ 0.1	87.3 $\pm$ 0.8/84.0 $\pm$ 0.3	
	GCN-EBM	91.6 $\pm$ 0.9/1.4 $\pm$ 0.3	90.3 $\pm$ 0.0	89.1 $\pm$ 0.8/9.0 $\pm$ 0.3	87.4 $\pm$ 0.1	75.2 $\pm$ 0.7/74.4 $\pm$ 0.3	75.8 $\pm$ 0.3	76.4 $\pm$ 0.7/75.5 $\pm$ 0.3	75.8 $\pm$ 0.3	46.8 $\pm$ 0.4/44.0 $\pm$ 0.3	75.8 $\pm$ 0.3	80.4 $\pm$ 0.7/83.1 $\pm$ 0.3	82.6 $\pm$ 0.1	86.0 $\pm$ 0.8/84.1 $\pm$ 0.3	
	GCN-HEAT	91.6 $\pm$ 0.9/1.4 $\pm$ 0.3	90.3 $\pm$ 0.0	89.1 $\pm$ 0.8/9.0 $\pm$ 0.3	87.4 $\pm$ 0.1	75.2 $\pm$ 0.7/74.4 $\pm$ 0.3	75.8 $\pm$ 0.3	76.4 $\pm$ 0.7/75.5 $\pm$ 0.3	75.8 $\pm$ 0.3	46.8 $\pm$ 0.4/44.0 $\pm$ 0.3	75.8 $\pm$ 0.3	80.4 $\pm$ 0.7/83.1 $\pm$ 0.3	82.6 $\pm$ 0.1	86.0 $\pm$ 0.8/84.1 $\pm$ 0.3	
	GCnSafe	91.6 $\pm$ 0.9/1.4 $\pm$ 0.3	90.3 $\pm$ 0.0	89.1 $\pm$ 0.8/9.0 $\pm$ 0.3	87.4 $\pm$ 0.1	75.2 $\pm$ 0.7/74.4 $\pm$ 0.3	75.8 $\pm$ 0.3	76.4 $\pm$ 0.7/75.5 $\pm$ 0.3	75.8 $\pm$ 0.3	46.8 $\pm$ 0.4/44.0 $\pm$ 0.3	75.8 $\pm$ 0.3	80.4 $\pm$ 0.7/83.1 $\pm$ 0.3	82.6 $\pm$ 0.1	86.0 $\pm$ 0.8/84.1 $\pm$ 0.3	
	GCN-GERM	91.6 $\pm$ 0.9/1.4 $\pm$ 0.3	90.3 $\pm$ 0.0	89.1 $\pm$ 0.8/9.0 $\pm$ 0.3	87.4 $\pm$ 0.1	75.2 $\pm$ 0.7/74.4 $\pm$ 0.3	75.8 $\pm$ 0.3	76.4 $\pm$ 0.7/75.5 $\pm$ 0.3	75.8 $\pm$ 0.3	46.8 $\pm$ 0.4/44.0 $\pm$ 0.3	75.8 $\pm$ 0.3	80.4 $\pm$ 0.7/83.1 $\pm$ 0.3	82.6 $\pm$ 0.1	86.0 $\pm$ 0.8/84.1 $\pm$ 0.3	
CoraML-LLM	GCN-MCD	90.0 $\pm$ 0.5/0.0 $\pm$ 0.3	90.1 $\pm$ 0.0	89.2 $\pm$ 0.4/49.0 $\pm$ 0.5	89.5 $\pm$ 0.3	n.a./n.a.	n.a.	77.5 $\pm$ 0.5/61.0 $\pm$ 0.3	76.9 $\pm$ 0.1	51.7 $\pm$ 0.5/50.1 $\pm$ 0.3	76.9 $\pm$ 0.1	85.2 $\pm$ 0.5/60.3 $\pm$ 0.3	84.0 $\pm$ 0.1	87.2 $\pm$ 0.5/60.7 $\pm$ 0.3	91.4 $\pm$ 0.1
	GCN-DE	88.4 $\pm$ 0.7/6.0 $\pm$ 0.3	87.8 $\pm$ 0.1	87.4 $\pm$ 0.7/4.8 $\pm$ 0.3	87.3 $\pm$ 0.2	n.a./n.a.	n.a.	75.5 $\pm$ 0.5/54.4 $\pm$ 0.3	75.6 $\pm$ 0.1	48.7 $\pm$ 0.4/64.7 $\pm$ 0.3	75.6 $\pm$ 0.1	83.8 $\pm$ 0.7/64.3 $\pm$ 0.3	81.8 $\pm$ 0.2	84.2 $\pm$ 0.7/65.2 $\pm$ 0.3	88.7 $\pm$ 0.1
	GCN-MCD+DE	88.5 $\pm$ 0.7/6.0 $\pm$ 0.3	87.9 $\pm$ 0.1	87.5 $\pm$ 0.7/5.2 $\pm$ 0.3	87.1 $\pm$ 0.1	n.a./n.a.	n.a.	75.7 $\pm$ 0.5/54.4 $\pm$ 0.3	75.7 $\pm$ 0.1	48.7 $\pm$ 0.4/64.7 $\pm$ 0.3	75.7 $\pm$ 0.1	83.9 $\pm$ 0.7/65.6 $\pm$ 0.3	81.9 $\pm$ 0.2	84.2 $\pm$ 0.7/65.2 $\pm$ 0.3	88.7 $\pm$ 0.1
	GCN-BNN	90.8 $\pm$ 0.8/5.7 $\pm$ 0.3	90.0 $\pm$ 0.1	88.8 $\pm$ 0.6/86.2 $\pm$ 0.3	89.3 $\pm$ 0.1	n.a./n.a.	n.a.	77.9 $\pm$ 0.5/71.1 $\pm$ 0.3	76.8 $\pm$ 0.1	54.0 $\pm$ 0.5/67.6 $\pm$ 0.3	76.8 $\pm$ 0.1	85.7 $\pm$ 0.8/83.9 $\pm$ 0.3	83.7 $\pm$ 0.1	88.5 $\pm$ 0.8/86.1 $\pm$ 0.3	91.4 $\pm$ 0.1
	GCN-Ens	90.0 $\pm$ 0.7/9.1 $\pm$ 0.3	90.1 $\pm$ 0.0	89.2 $\pm$ 0.7/9.5 $\pm$ 0.3	89.5 $\pm$ 0.1	n.a./n.a.	n.a.	77.4 $\pm$ 0.5/74.2 $\pm$ 0.3	77.3 $\pm$ 0.1	52.2 $\pm$ 0.5/67.0 $\pm$ 0.3	77.3 $\pm$ 0.1	85.2 $\pm$ 0.7/73.4 $\pm$ 0.3	84.2 $\pm$ 0.1	87.1 $\pm$ 0.7/72.9 $\pm$ 0.3	91.4 $\pm$ 0.1
	CPN	89.6 $\pm$ 0.8/6.0 $\pm$ 0.3	89.0 $\pm$ 0.1	88.5 $\pm$ 0.8/85.1 $\pm$ 0.3	89.1 $\pm$ 0.1	n.a./n.a.	n.a.	82.5 $\pm$ 0.5/71.5 $\pm$ 0.3	75.8 $\pm$ 0.1	43.6 $\pm$ 0.5/77.3 $\pm$ 0.3	75.8 $\pm$ 0.1	84.6 $\pm$ 0.7/78.9 $\pm$ 0.3	81.3 $\pm$ 0.1	82.5 $\pm$ 0.7/72.4 $\pm$ 0.3	90.8 $\pm$ 0.1
	SGCN	90.4 $\pm$ 0.8/7.6 $\pm$ 0.3	88.9 $\pm$ 0.0	90.4 $\pm$ 0.8/8.2 $\pm$ 0.3	90.4 $\pm$ 0.1	n.a./n.a.	n.a.	78.5 $\pm$ 0.5/61.7 $\pm$ 0.3	68.1 $\pm$ 0.1	51.5 $\pm$ 0.5/48.8 $\pm$ 0.3	68.1 $\pm$ 0.1	84.2 $\pm$ 0.7/73.0 $\pm$ 0.3	78.9 $\pm$ 0.1	82.7 $\pm$ 0.7/73.0 $\pm$ 0.3	90.2 $\pm$ 0.1
	GCN-EBM	89.9 $\pm$ 0.8/9.2 $\pm$ 0.3	90.2 $\pm$ 0.0	89.1 $\pm$ 0.8/88.5 $\pm$ 0.3	89.4 $\pm$ 0.1	n.a./n.a.	n.a.	78.1 $\pm$ 0.5/76.0 $\pm$ 0.3	77.1 $\pm$ 0.1	51.3 $\pm$ 0.5/48.3 $\pm$ 0.3	77.1 $\pm$ 0.1	85.2 $\pm$ 0.8/80.9 $\pm$ 0.3	83.7 $\pm$ 0.1	87.1 $\pm$ 0.8/84.9 $\pm$ 0.3	91.1 $\pm$ 0.1
	GCN-HEAT	89.9 $\pm$ 0.8/9.2 $\pm$ 0.3	90.2 $\pm$ 0.0	89.1 $\pm$ 0.8/88.5 $\pm$ 0.3	89.4 $\pm$ 0.1	n.a./n.a.	n.a.	78.1 $\pm$ 0.5/76.0 $\pm$ 0.3	77.1 $\pm$ 0.1	51.3 $\pm$ 0.5/48.3 $\pm$ 0.3	77.1 $\pm$ 0.1	85.2 $\pm$ 0.8/80.9 $\pm$ 0.3	83.7 $\pm$ 0.1	87.1 $\pm$ 0.8/84.9 $\pm$ 0.3	91.1 $\pm$ 0.1
	GCnSafe	89.9 $\pm$ 0.8/9.2 $\pm$ 0.3	90.2 $\pm$ 0.0	89.1 $\pm$ 0.8/88.5 $\pm$ 0.3	89.4 $\pm$ 0.1	n.a./n.a.	n.a.	78.1 $\pm$ 0.5/76.0 $\pm$ 0.3	77.1 $\pm$ 0.1	51.3 $\pm$ 0.5/48.3 $\pm$ 0.3	77.1 $\pm$ 0.1	85.2 $\pm$ 0.8/80.9 $\pm$ 0.3	83.7 $\pm$ 0.1	87.1 $\pm$ 0.8/84.9 $\pm$ 0.3	91.1 $\pm$ 0.1
	GCN-GERM	89.9 $\pm$ 0.8/9.2 $\pm$ 0.3	90.2 $\pm$ 0.0	89.1 $\pm$ 0.8/88.5 $\pm$ 0.3	89.4 $\pm$ 0.1	n.a./n.a.	n.a.	78.1 $\pm$ 0.5/76.0 $\pm$ 0.3	77.1 $\pm$ 0.1	51.3 $\pm$ 0.5/48.3 $\pm$ 0.3	77.1 $\pm$ 0.1	85.2 $\pm$ 0.8/80.9 $\pm$ 0.3	83.7 $\pm$ 0.1	87.1 $\pm$ 0.8/84.9 $\pm$ 0.3	91.1 $\pm$ 0.1
Citeseer	GCN-MCD	89.5 $\pm$ 0.5/0.2 $\pm$ 0.3	84.8 $\pm$ 0.1	90.5 $\pm$ 0.5/50.2 $\pm$ 0.3	90.4 $\pm$ 0.1	73.7 $\pm$ 0.5/50.4 $\pm$ 0.7	76.4 $\pm$ 0.1	72.6 $\pm$ 0.5/49.8 $\pm$ 0.3	76.4 $\pm$ 0.1	43.1 $\pm$ 0.5/50.5 $\pm$ 0.3	76.4 $\pm$ 0.1	83.4 $\pm$ 0.6/50.1 $\pm$ 0.3	78.4 $\pm$ 0.1	82.6 $\pm$ 0.5/50.4 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCN-DE	84.8 $\pm$ 0.7/6.0 $\pm$ 0.3	81.0 $\pm$ 0.1	87.1 $\pm$ 0.7/4.8 $\pm$ 0.3	86.4 $\pm$ 0.1	68.2 $\pm$ 0.5/70.7 $\pm$ 0.3	72.9 $\pm$ 0.1	62.7 $\pm$ 0.5/78.3 $\pm$ 0.3	72.9 $\pm$ 0.1	39.3 $\pm$ 0.5/61.9 $\pm$ 0.3	72.9 $\pm$ 0.1	80.0 $\pm$ 0.7/63.1 $\pm$ 0.3	75.6 $\pm$ 0.1	80.8 $\pm$ 0.7/65.2 $\pm$ 0.3	87.3 $\pm$ 0.1
	GCN-MCD+DE	85.1 $\pm$ 0.7/6.0 $\pm$ 0.3	81.3 $\pm$ 0.1	87.5 $\pm$ 0.7/5.2 $\pm$ 0.3	87.1 $\pm$ 0.1	66.8 $\pm$ 0.5/70.4 $\pm$ 0.3	73.0 $\pm$ 0.1	63.4 $\pm$ 0.5/77.1 $\pm$ 0.3	73.0 $\pm$ 0.1	39.6 $\pm$ 0.5/63.9 $\pm$ 0.3	73.0 $\pm$ 0.1	80.1 $\pm$ 0.7/65.6 $\pm$ 0.3	75.0 $\pm$ 0.1	81.0 $\pm$ 0.7/65.2 $\pm$ 0.3	87.3 $\pm$ 0.1
	GCN-BNN	89.2 $\pm$ 0.6/1.1 $\pm$ 0.3	84.5 $\pm$ 0.1	90.1 $\pm$ 0.8/9.0 $\pm$ 0.3	90.3 $\pm$ 0.1	75.8 $\pm$ 0.7/72.1 $\pm$ 0.3	75.9 $\pm$ 0.1	68.6 $\pm$ 0.7/78.6 $\pm$ 0.3	75.9 $\pm$ 0.1	43.5 $\pm$ 0.5/59.7 $\pm$ 0.3	75.9 $\pm$ 0.1	84.2 $\pm$ 0.7/69.0 $\pm$ 0.3	77.8 $\pm$ 0.1	84.4 $\pm$ 0.8/85.1 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCN-Ens	89.6 $\pm$ 0.8/1.0 $\pm$ 0.3	84.8 $\pm$ 0.1	90.4 $\pm$ 0.8/1.3 $\pm$ 0.3	91.1 $\pm$ 0.1	74.2 $\pm$ 0.5/58.6 $\pm$ 0.3	76.9 $\pm$ 0.1	73.2 $\pm$ 0.5/74.9 $\pm$ 0.3	76.9 $\pm$ 0.1	44.4 $\pm$ 0.5/58.9 $\pm$ 0.3	76.9 $\pm$ 0.1	83.2 $\pm$ 0.7/79.3 $\pm$ 0.3	78.6 $\pm$ 0.1	82.6 $\pm$ 0.7/76.2 $\pm$ 0.3	90.5 $\pm$ 0.1
	CPN	89.6 $\pm$ 0.8/6.0 $\pm$ 0.3	85.6 $\pm$ 0.1	92.6 $\pm$ 0.8/87.3 $\pm$ 0.3	91.6 $\pm$ 0.1	81.8 $\pm$ 0.7/77.4 $\pm$ 0.3	76.7 $\pm$ 0.1	82.7 $\pm$ 0.5/82.9 $\pm$ 0.3	76.7 $\pm$ 0.1	46.8 $\pm$ 0.5/80.1 $\pm$ 0.3	76.7 $\pm$ 0.1	86.8 $\pm$ 0.8/81.1 $\pm$ 0.3	78.3 $\pm$ 0.1	87.4 $\pm$ 0.8/79.9 $\pm$ 0.3	91.2 $\pm$ 0.1
	SGCN	89.4 $\pm$ 0.8/1.2 $\pm$ 0.3	85.0 $\pm$ 0.1	90.4 $\pm$ 0.8/8.2 $\pm$ 0.3	90.4 $\pm$ 0.1	75.9 $\pm$ 0.7/73.5 $\pm$ 0.3	75.8 $\pm$ 0.1	79.6 $\pm$ 0.7/77.4 $\pm$ 0.3	75.8 $\pm$ 0.1	43.9 $\pm$ 0.5/41.7 $\pm$ 0.3	75.8 $\pm$ 0.1	83.6 $\pm$ 0.8/82.1 $\pm$ 0.3	79.3 $\pm$ 0.1	83.6 $\pm$ 0.8/82.1 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCN-EBM	89.5 $\pm$ 0.8/9.5 $\pm$ 0.3	84.8 $\pm$ 0.0	90.5 $\pm$ 0.8/1.0 $\pm$ 0.3	90.9 $\pm$ 0.1	75.2 $\pm$ 0.7/74.6 $\pm$ 0.3	76.6 $\pm$ 0.1	73.9 $\pm$ 0.7/76.6 $\pm$ 0.3	76.6 $\pm$ 0.1	45.3 $\pm$ 0.5/40.9 $\pm$ 0.3	76.6 $\pm$ 0.1	83.2 $\pm$ 0.8/81.1 $\pm$ 0.3	78.7 $\pm$ 0.1	82.8 $\pm$ 0.8/81.1 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCN-HEAT	89.5 $\pm$ 0.8/9.5 $\pm$ 0.3	84.8 $\pm$ 0.0	90.5 $\pm$ 0.8/1.0 $\pm$ 0.3	90.9 $\pm$ 0.1	75.2 $\pm$ 0.7/74.6 $\pm$ 0.3	76.6 $\pm$ 0.1	73.9 $\pm$ 0.7/76.6 $\pm$ 0.3	76.6 $\pm$ 0.1	45.3 $\pm$ 0.5/40.9 $\pm$ 0.3	76.6 $\pm$ 0.1	83.2 $\pm$ 0.8/81.1 $\pm$ 0.3	78.7 $\pm$ 0.1	82.8 $\pm$ 0.8/81.1 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCnSafe	89.5 $\pm$ 0.8/9.5 $\pm$ 0.3	84.8 $\pm$ 0.0	90.5 $\pm$ 0.8/1.0 $\pm$ 0.3	90.9 $\pm$ 0.1	75.2 $\pm$ 0.7/74.6 $\pm$ 0.3	76.6 $\pm$ 0.1	73.9 $\pm$ 0.7/76.6 $\pm$ 0.3	76.6 $\pm$ 0.1	45.3 $\pm$ 0.5/40.9 $\pm$ 0.3	76.6 $\pm$ 0.1	83.2 $\pm$ 0.8/81.1 $\pm$ 0.3	78.7 $\pm$ 0.1	82.8 $\pm$ 0.8/81.1 $\pm$ 0.3	91.2 $\pm$ 0.1
	GCN-GERM	89.5 $\pm$ 0.8/9.5 $\pm$ 0.3	84.8 $\pm$ 0.0	90.5 $\pm$ 0.8/1.0 $\pm$ 0.3	90.9 $\pm$ 0.1	75.2 $\pm$ 0.7/74.6 $\pm$ 0.3	76.6 $\pm$ 0.1	73.9 $\pm$ 0.7/76.6 $\pm$ 0.3	76.6 $\pm$ 0.1	45.3 $\pm$ 0.5/40.9 $\pm$ 0.3	76.6 $\pm$ 0.1	83.2 $\pm$ 0.8/81.1 $\pm</$			

	Model	LoC (best)		LoC (runner-up)		Bert(0.1) (best)		Bert(0.1) (runner-up)		Page Rank		Homophily			
		AUC-ROC↑ (Area, F1s)	Acc.↑	AUC-ROC↑ (Area, F1s)	Acc.↑	AUC-ROC↑ (Area, F1s)	Acc.↑	AUC-ROC↑ (Area, F1s)	Acc.↑	AUC-ROC↑ (Area, F1s)	Acc.↑	AUC-ROC↑ (Area, F1s)	Acc.↑		
CoAtLLM	GCM-MCD	89.6 _{1.1} /70.1 _{1.1}	90.4 _{1.1}	86.3 _{1.1} /69.1 _{1.1}	85.2 _{1.1}	79.1 _{1.1} /59.0 _{1.1}	82.0 _{1.1}	77.6 _{1.1} /50.1 _{1.1}	82.0 _{1.1}	47.8 _{1.1} /49.9 _{1.1}	82.0 _{1.1}	79.1 _{1.1} /49.5 _{1.1}	82.6 _{1.1}	86.0 _{1.1} /49.8 _{1.1}	93.9 _{1.1}
	GCM-DE	87.2 _{1.1} /75.6 _{1.1}	87.3 _{1.1}	83.8 _{1.1} /71.3 _{1.1}	80.9 _{1.1}	70.7 _{1.1} /71.5 _{1.1}	78.4 _{1.1}	66.3 _{1.1} /71.5 _{1.1}	78.4 _{1.1}	42.1 _{1.1} /46.6 _{1.1}	78.4 _{1.1}	77.0 _{1.1} /68.9 _{1.1}	80.6 _{1.1}	87.1 _{1.1} /70.0 _{1.1}	91.4 _{1.1}
	GCM-MCD+DE	87.5 _{1.1} /76.6 _{1.1}	87.6 _{1.1}	83.3 _{1.1} /71.9 _{1.1}	80.1 _{1.1}	69.9 _{1.1} /71.4 _{1.1}	78.1 _{1.1}	66.0 _{1.1} /72.0 _{1.1}	78.1 _{1.1}	42.5 _{1.1} /46.8 _{1.1}	78.1 _{1.1}	76.7 _{1.1} /70.6 _{1.1}	80.5 _{1.1}	82.9 _{1.1} /75.3 _{1.1}	91.1 _{1.1}
	GCM-BNN	88.0 _{1.1} /88.1 _{1.1}	89.0 _{1.1}	83.3 _{1.1} /82.6 _{1.1}	82.1 _{1.1}	79.4 _{1.1} /76.0 _{1.1}	80.6 _{1.1}	77.3 _{1.1} /78.4 _{1.1}	81.8 _{1.1}	50.2 _{1.1} /55.2 _{1.1}	80.6 _{1.1}	78.9 _{1.1} /77.1 _{1.1}	80.4 _{1.1}	85.8 _{1.1} /85.3 _{1.1}	93.1 _{1.1}
	GCM-Ens	89.8 _{1.1} /87.3 _{1.1}	90.3 _{1.1}	85.5 _{1.1} /82.5 _{1.1}	85.2 _{1.1}	78.1 _{1.1} /81.9 _{1.1}	81.8 _{1.1}	77.3 _{1.1} /85.6 _{1.1}	81.8 _{1.1}	48.3 _{1.1} /46.6 _{1.1}	81.8 _{1.1}	78.9 _{1.1} /80.0 _{1.1}	82.7 _{1.1}	88.1 _{1.1} /87.6 _{1.1}	93.6 _{1.1}
	GPIN	88.9 _{1.1} /83.9 _{1.1}	90.0 _{1.1}	85.3 _{1.1} /82.7 _{1.1}	85.5 _{1.1}	82.2 _{1.1} /86.7 _{1.1}	78.8 _{1.1}	80.5 _{1.1} /87.6 _{1.1}	78.8 _{1.1}	82.1 _{1.1} /87.3 _{1.1}	78.8 _{1.1}	81.6 _{1.1} /88.7 _{1.1}	78.4 _{1.1}	83.8 _{1.1} /87.3 _{1.1}	91.1 _{1.1}
	SGCN	89.8 _{1.1} /89.9 _{1.1}	91.0 _{1.1}	88.0 _{1.1} /85.9 _{1.1}	88.5 _{1.1}	72.8 _{1.1} /80.2 _{1.1}	81.7 _{1.1}	68.1 _{1.1} /83.1 _{1.1}	81.7 _{1.1}	49.2 _{1.1} /47.3 _{1.1}	81.7 _{1.1}	82.8 _{1.1} /87.7 _{1.1}	83.3 _{1.1}	86.0 _{1.1} /83.2 _{1.1}	94.2 _{1.1}
	GCM-EBM	89.7 _{1.1} /89.2 _{1.1}	90.4 _{1.1}	86.4 _{1.1} /83.3 _{1.1}	85.0 _{1.1}	77.4 _{1.1} /74.9 _{1.1}	81.9 _{1.1}	78.2 _{1.1} /74.4 _{1.1}	81.9 _{1.1}	45.6 _{1.1} /46.2 _{1.1}	81.9 _{1.1}	79.2 _{1.1} /77.8 _{1.1}	82.5 _{1.1}	85.8 _{1.1} /85.3 _{1.1}	94.1 _{1.1}
	GCM-HEAT	89.7 _{1.1} /87.9 _{1.1}	90.4 _{1.1}	86.4 _{1.1} /83.3 _{1.1}	85.0 _{1.1}	77.4 _{1.1} /74.9 _{1.1}	81.9 _{1.1}	78.2 _{1.1} /74.4 _{1.1}	81.9 _{1.1}	45.6 _{1.1} /46.2 _{1.1}	81.9 _{1.1}	79.2 _{1.1} /77.8 _{1.1}	82.5 _{1.1}	85.8 _{1.1} /84.0 _{1.1}	94.1 _{1.1}
	GCMsE	89.7 _{1.1} /91.7 _{1.1}	90.4 _{1.1}	86.4 _{1.1} /88.4 _{1.1}	85.0 _{1.1}	77.4 _{1.1} /74.9 _{1.1}	81.9 _{1.1}	78.2 _{1.1} /77.1 _{1.1}	81.9 _{1.1}	45.6 _{1.1} /43.1 _{1.1}	81.9 _{1.1}	79.2 _{1.1} /77.8 _{1.1}	82.5 _{1.1}	85.8 _{1.1} /84.0 _{1.1}	94.1 _{1.1}
CoAtLLM	GCM-MCD	89.7 _{1.1} /91.1 _{1.1}	90.4 _{1.1}	86.4 _{1.1} /85.7 _{1.1}	85.0 _{1.1}	77.4 _{1.1} /74.9 _{1.1}	81.9 _{1.1}	78.2 _{1.1} /77.1 _{1.1}	81.9 _{1.1}	45.6 _{1.1} /57.7 _{1.1}	81.9 _{1.1}	79.2 _{1.1} /77.8 _{1.1}	82.5 _{1.1}	85.8 _{1.1} /83.2 _{1.1}	94.1 _{1.1}
	GCM-DE	90.9 _{1.1} /90.5 _{1.1}	91.2 _{1.1}	89.0 _{1.1} /86.1 _{1.1}	90.6 _{1.1}	n.a.	n.a.	79.4 _{1.1} /49.4 _{1.1}	82.6 _{1.1}	52.3 _{1.1} /50.0 _{1.1}	82.6 _{1.1}	84.9 _{1.1} /49.5 _{1.1}	84.9 _{1.1}	88.5 _{1.1} /50.1 _{1.1}	94.1 _{1.1}
	GCM-MCD+DE	89.6 _{1.1} /80.3 _{1.1}	89.7 _{1.1}	87.6 _{1.1} /76.9 _{1.1}	88.9 _{1.1}	n.a.	n.a.	70.6 _{1.1} /59.2 _{1.1}	79.9 _{1.1}	53.3 _{1.1} /48.0 _{1.1}	80.3 _{1.1}	82.7 _{1.1} /74.8 _{1.1}	82.2 _{1.1}	86.1 _{1.1} /74.8 _{1.1}	91.4 _{1.1}
	GCM-BNN	91.0 _{1.1} /88.5 _{1.1}	91.4 _{1.1}	88.5 _{1.1} /86.1 _{1.1}	91.0 _{1.1}	n.a.	n.a.	70.1 _{1.1} /68.2 _{1.1}	82.8 _{1.1}	55.5 _{1.1} /49.4 _{1.1}	82.8 _{1.1}	85.3 _{1.1} /83.0 _{1.1}	84.0 _{1.1}	89.2 _{1.1} /87.7 _{1.1}	94.0 _{1.1}
	GCM-Ens	90.8 _{1.1} /86.8 _{1.1}	91.4 _{1.1}	89.1 _{1.1} /87.0 _{1.1}	90.6 _{1.1}	n.a.	n.a.	72.1 _{1.1} /55.2 _{1.1}	82.6 _{1.1}	53.8 _{1.1} /48.0 _{1.1}	82.6 _{1.1}	84.7 _{1.1} /73.6 _{1.1}	85.0 _{1.1}	88.1 _{1.1} /87.6 _{1.1}	93.6 _{1.1}
	GPIN	90.0 _{1.1} /91.2 _{1.1}	90.5 _{1.1}	88.0 _{1.1} /85.9 _{1.1}	88.5 _{1.1}	n.a.	n.a.	82.9 _{1.1} /73.6 _{1.1}	79.7 _{1.1}	82.7 _{1.1} /73.5 _{1.1}	79.7 _{1.1}	84.5 _{1.1} /74.9 _{1.1}	80.9 _{1.1}	85.3 _{1.1} /77.9 _{1.1}	92.5 _{1.1}
	SGCN	91.7 _{1.1} /90.7 _{1.1}	90.3 _{1.1}	89.6 _{1.1} /85.9 _{1.1}	90.6 _{1.1}	n.a.	n.a.	78.8 _{1.1} /82.2 _{1.1}	68.7 _{1.1}	51.4 _{1.1} /48.3 _{1.1}	68.7 _{1.1}	84.6 _{1.1} /74.9 _{1.1}	82.0 _{1.1}	87.0 _{1.1} /82.8 _{1.1}	93.1 _{1.1}
	GCM-EBM	90.7 _{1.1} /90.0 _{1.1}	91.4 _{1.1}	89.1 _{1.1} /87.9 _{1.1}	90.7 _{1.1}	n.a.	n.a.	78.8 _{1.1} /76.6 _{1.1}	82.6 _{1.1}	55.2 _{1.1} /50.3 _{1.1}	82.6 _{1.1}	84.8 _{1.1} /81.0 _{1.1}	85.0 _{1.1}	88.1 _{1.1} /87.6 _{1.1}	93.6 _{1.1}
	GCM-HEAT	90.7 _{1.1} /86.3 _{1.1}	91.4 _{1.1}	89.1 _{1.1} /84.7 _{1.1}	90.7 _{1.1}	n.a.	n.a.	78.8 _{1.1} /73.5 _{1.1}	82.6 _{1.1}	55.2 _{1.1} /46.1 _{1.1}	82.6 _{1.1}	84.8 _{1.1} /75.8 _{1.1}	85.0 _{1.1}	88.6 _{1.1} /83.0 _{1.1}	93.8 _{1.1}
	GCMsE	90.7 _{1.1} /92.6 _{1.1}	91.4 _{1.1}	89.1 _{1.1} /91.4 _{1.1}	90.7 _{1.1}	n.a.	n.a.	78.8 _{1.1} /80.3 _{1.1}	82.6 _{1.1}	55.2 _{1.1} /52.8 _{1.1}	82.6 _{1.1}	84.8 _{1.1} /83.1 _{1.1}	85.0 _{1.1}	88.6 _{1.1} /86.2 _{1.1}	94.1 _{1.1}
Amazon Computers	GCM-MCD	88.0 _{1.1} /49.2 _{1.1}	83.5 _{1.1}	88.6 _{1.1} /49.3 _{1.1}	91.1 _{1.1}	74.4 _{1.1} /50.1 _{1.1}	82.1 _{1.1}	77.1 _{1.1} /50.6 _{1.1}	82.1 _{1.1}	45.4 _{1.1} /49.5 _{1.1}	82.1 _{1.1}	82.6 _{1.1} /50.3 _{1.1}	82.9 _{1.1}	84.6 _{1.1} /50.2 _{1.1}	96.3 _{1.1}
	GCM-DE	85.2 _{1.1} /77.6 _{1.1}	80.1 _{1.1}	85.9 _{1.1} /74.4 _{1.1}	86.4 _{1.1}	76.8 _{1.1}	76.8 _{1.1}	64.8 _{1.1} /79.0 _{1.1}	76.8 _{1.1}	41.4 _{1.1} /46.1 _{1.1}	76.8 _{1.1}	81.9 _{1.1} /79.7 _{1.1}	81.6 _{1.1}	89.1 _{1.1} /79.7 _{1.1}	81.6 _{1.1}
	GCM-MCD+DE	84.9 _{1.1} /77.2 _{1.1}	80.2 _{1.1}	85.5 _{1.1} /75.5 _{1.1}	86.2 _{1.1}	68.4 _{1.1} /73.6 _{1.1}	76.8 _{1.1}	65.5 _{1.1} /79.4 _{1.1}	76.8 _{1.1}	40.4 _{1.1} /46.1 _{1.1}	76.8 _{1.1}	80.0 _{1.1} /76.6 _{1.1}	82.0 _{1.1}	82.3 _{1.1} /79.9 _{1.1}	81.7 _{1.1}
	GCM-BNN	88.0 _{1.1} /88.2 _{1.1}	89.2 _{1.1}	88.4 _{1.1} /87.8 _{1.1}	91.1 _{1.1}	74.2 _{1.1} /77.4 _{1.1}	82.0 _{1.1}	72.5 _{1.1} /83.2 _{1.1}	82.0 _{1.1}	47.8 _{1.1} /49.9 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /83.2 _{1.1}	82.7 _{1.1}	84.9 _{1.1} /85.6 _{1.1}	86.7 _{1.1}
	GCM-Ens	89.7 _{1.1} /89.8 _{1.1}	91.1 _{1.1}	88.7 _{1.1} /81.1 _{1.1}	91.2 _{1.1}	73.7 _{1.1} /77.0 _{1.1}	82.0 _{1.1}	70.2 _{1.1} /77.2 _{1.1}	82.0 _{1.1}	46.5 _{1.1} /45.8 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /75.9 _{1.1}	82.6 _{1.1}	86.1 _{1.1} /85.6 _{1.1}	93.1 _{1.1}
	GPIN	88.7 _{1.1} /87.3 _{1.1}	84.9 _{1.1}	91.2 _{1.1} /86.4 _{1.1}	90.9 _{1.1}	83.5 _{1.1} /80.9 _{1.1}	83.3 _{1.1}	83.1 _{1.1} /85.9 _{1.1}	83.3 _{1.1}	88.1 _{1.1} /82.9 _{1.1}	83.3 _{1.1}	86.3 _{1.1} /78.6 _{1.1}	82.1 _{1.1}	88.0 _{1.1} /81.9 _{1.1}	92.5 _{1.1}
	SGCN	88.0 _{1.1} /89.7 _{1.1}	83.5 _{1.1}	85.1 _{1.1} /80.8 _{1.1}	90.6 _{1.1}	76.0 _{1.1} /73.3 _{1.1}	81.5 _{1.1}	81.3 _{1.1} /79.0 _{1.1}	82.0 _{1.1}	45.5 _{1.1} /44.1 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /80.0 _{1.1}	82.5 _{1.1}	85.0 _{1.1} /83.5 _{1.1}	95.9 _{1.1}
	GCM-EBM	87.0 _{1.1} /87.9 _{1.1}	83.6 _{1.1}	88.4 _{1.1} /86.0 _{1.1}	91.1 _{1.1}	74.2 _{1.1} /77.4 _{1.1}	82.0 _{1.1}	70.4 _{1.1} /70.5 _{1.1}	82.0 _{1.1}	45.6 _{1.1} /49.9 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /81.3 _{1.1}	82.4 _{1.1}	84.4 _{1.1} /82.4 _{1.1}	96.4 _{1.1}
	GCM-HEAT	87.0 _{1.1} /86.0 _{1.1}	83.6 _{1.1}	88.4 _{1.1} /86.0 _{1.1}	91.1 _{1.1}	74.2 _{1.1} /77.4 _{1.1}	82.0 _{1.1}	70.4 _{1.1} /74.3 _{1.1}	82.0 _{1.1}	45.6 _{1.1} /45.4 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /84.9 _{1.1}	82.4 _{1.1}	84.4 _{1.1} /85.9 _{1.1}	96.4 _{1.1}
	GCMsE	87.0 _{1.1} /86.0 _{1.1}	83.6 _{1.1}	88.4 _{1.1} /86.0 _{1.1}	91.1 _{1.1}	74.2 _{1.1} /77.4 _{1.1}	82.0 _{1.1}	70.4 _{1.1} /73.6 _{1.1}	82.0 _{1.1}	45.6 _{1.1} /44.0 _{1.1}	82.0 _{1.1}	82.7 _{1.1} /84.9 _{1.1}	82.4 _{1.1}	84.4 _{1.1} /85.9 _{1.1}	96.4 _{1.1}
Amazon Computers	GCM-MCD	71.4 _{1.1} /50.1 _{1.1}	91.6 _{1.1}	71.2 _{1.1} /50.1 _{1.1}	83.7 _{1.1}	n.a.	n.a.	70.3 _{1.1} /50.2 _{1.1}	78.1 _{1.1}	50.4 _{1.1} /49.8 _{1.1}	78.1 _{1.1}	73.2 _{1.1} /50.0 _{1.1}	77.9 _{1.1}	75.6 _{1.1} /49.9 _{1.1}	88.9 _{1.1}
	GCM-DE	70.4 _{1.1} /62.3 _{1.1}	91.2 _{1.1}	70.6 _{1.1} /65.7 _{1.1}	80.5 _{1.1}	n.a.	n.a.	64.0 _{1.1} /66.7 _{1.1}	74.9 _{1.1}	47.4 _{1.1} /62.7 _{1.1}	74.9 _{1.1}	70.2 _{1.1} /64.2 _{1.1}	76.0 _{1.1}	73.0 _{1.1} /68.6 _{1.1}	84.8 _{1.1}
	GCM-MCD+DE	71.2 _{1.1} /62.3 _{1.1}	90.8 _{1.1}	71.0 _{1.1} /66.0 _{1.1}	80.7 _{1.1}	n.a.	n.a.	64.0 _{1.1} /67.4 _{1.1}	75.4 _{1.1}	47.0 _{1.1} /62.5 _{1.1}	75.4 _{1.1}	70.5 _{1.1} /65.0 _{1.1}	76.3 _{1.1}	73.1 _{1.1} /68.9 _{1.1}	84.9 _{1.1}
	GCM-BNN	71.3 _{1.1} /71.2 _{1.1}	92.1 _{1.1}	72.6 _{1.1} /72.1 _{1.1}	85.1 _{1.1}	n.a.	n.a.	70.9 _{1.1} /69.2 _{1.1}	80.1 _{1.1}	50.9 _{1.1} /47.8 _{1.1}	80.1 _{1.1}	75.1 _{1.1} /73.8 _{1.1}	76.1 _{1.1}	76.1 _{1.1} /68.9 _{1.1}	88.8 _{1.1}
	GCM-Ens	71.4 _{1.1} /54.2 _{1.1}													

	Model	LoC (best)		LoC (heuristics)		Ber(β) (near)		Ber(0.5)		A(0,1) (far)		Page Rank		Homophily	
		AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$	AUC-PR $\uparrow$	Acc $\uparrow$
CoraML	GCN-MCD	90.5 $\pm$ 0.51.5 $\pm$ 1.5	90.4 $\pm$ 0.5	84.7 $\pm$ 0.42.1 $\pm$ 0.4	87.4 $\pm$ 0.5	51.8 $\pm$ 0.29.3 $\pm$ 0.3	75.6 $\pm$ 0.3	54.0 $\pm$ 0.23.8 $\pm$ 0.3	75.6 $\pm$ 0.3	74.5 $\pm$ 0.76.7 $\pm$ 0.7	75.6 $\pm$ 0.3	43.1 $\pm$ 0.20.7 $\pm$ 0.2	82.3 $\pm$ 0.5	52.3 $\pm$ 0.23.7 $\pm$ 0.3	91.4 $\pm$ 0.5
	GCN-DE	87.3 $\pm$ 0.47.2 $\pm$ 0.5	87.2 $\pm$ 0.7	80.6 $\pm$ 0.60.8 $\pm$ 0.8	82.0 $\pm$ 0.5	50.6 $\pm$ 0.52.8 $\pm$ 0.5	71.9 $\pm$ 0.7	58.4 $\pm$ 0.65.8 $\pm$ 0.6	71.9 $\pm$ 0.7	70.9 $\pm$ 0.48.1 $\pm$ 0.4	71.9 $\pm$ 0.7	47.1 $\pm$ 0.38.4 $\pm$ 0.3	79.3 $\pm$ 0.5	51.6 $\pm$ 0.34.7 $\pm$ 0.3	85.6 $\pm$ 0.5
	GCN-MCD+DE	87.6 $\pm$ 0.47.2 $\pm$ 0.5	87.3 $\pm$ 0.5	81.1 $\pm$ 0.61.0 $\pm$ 0.3	83.0 $\pm$ 0.5	50.6 $\pm$ 0.53.1 $\pm$ 0.5	72.1 $\pm$ 0.5	57.3 $\pm$ 0.64.6 $\pm$ 0.6	72.1 $\pm$ 0.5	72.2 $\pm$ 0.83.1 $\pm$ 0.4	72.1 $\pm$ 0.5	47.2 $\pm$ 0.37.9 $\pm$ 0.3	78.9 $\pm$ 0.5	51.2 $\pm$ 0.34.6 $\pm$ 0.3	85.9 $\pm$ 0.5
	GCN-BNN	88.4 $\pm$ 0.48.5 $\pm$ 0.4	89.0 $\pm$ 0.5	81.0 $\pm$ 0.67.1 $\pm$ 0.7	84.2 $\pm$ 0.5	54.9 $\pm$ 0.46.7 $\pm$ 0.5	75.1 $\pm$ 0.5	70.4 $\pm$ 0.69.3 $\pm$ 0.6	75.1 $\pm$ 0.5	78.5 $\pm$ 0.81.7 $\pm$ 0.9	73.1 $\pm$ 0.5	49.0 $\pm$ 0.47.1 $\pm$ 0.4	78.9 $\pm$ 0.5	56.8 $\pm$ 0.48.6 $\pm$ 0.4	88.6 $\pm$ 0.5
	GCN-Ess	90.7 $\pm$ 0.47.2 $\pm$ 0.5	90.4 $\pm$ 0.5	84.5 $\pm$ 0.47.5 $\pm$ 0.5	87.3 $\pm$ 0.5	47.2 $\pm$ 0.25.6 $\pm$ 0.3	75.1 $\pm$ 0.5	54.8 $\pm$ 0.55.9 $\pm$ 0.5	75.1 $\pm$ 0.5	74.4 $\pm$ 0.79.8 $\pm$ 0.7	75.1 $\pm$ 0.5	42.8 $\pm$ 0.27.2 $\pm$ 0.2	82.7 $\pm$ 0.5	55.2 $\pm$ 0.24.1 $\pm$ 0.2	91.5 $\pm$ 0.5
	GCN	87.2 $\pm$ 0.48.8 $\pm$ 0.5	88.5 $\pm$ 0.5	82.5 $\pm$ 0.47.4 $\pm$ 0.4	86.4 $\pm$ 0.5	56.5 $\pm$ 0.49.8 $\pm$ 0.5	72.8 $\pm$ 0.5	56.2 $\pm$ 0.41.4 $\pm$ 0.4	72.8 $\pm$ 0.5	59.0 $\pm$ 0.40.1 $\pm$ 0.1	72.8 $\pm$ 0.5	53.6 $\pm$ 0.37.1 $\pm$ 0.3	78.4 $\pm$ 0.5	59.5 $\pm$ 0.42.9 $\pm$ 0.4	87.6 $\pm$ 0.5
	SGCN	89.5 $\pm$ 0.48.7 $\pm$ 0.5	90.2 $\pm$ 0.5	86.5 $\pm$ 0.48.2 $\pm$ 0.5	89.9 $\pm$ 0.5	82.3 $\pm$ 0.47.1 $\pm$ 0.3	96.2 $\pm$ 0.5	83.0 $\pm$ 0.47.9 $\pm$ 0.6	96.2 $\pm$ 0.5	82.6 $\pm$ 0.48.3 $\pm$ 0.7	96.2 $\pm$ 0.5	52.1 $\pm$ 0.39.0 $\pm$ 0.3	74.8 $\pm$ 0.5	58.4 $\pm$ 0.38.9 $\pm$ 0.3	90.0 $\pm$ 0.5
	GCN-EBM	90.5 $\pm$ 0.49.0 $\pm$ 0.5	90.3 $\pm$ 0.5	84.5 $\pm$ 0.48.4 $\pm$ 0.4	87.4 $\pm$ 0.5	48.2 $\pm$ 0.47.1 $\pm$ 0.3	75.8 $\pm$ 0.5	54.5 $\pm$ 0.65.7 $\pm$ 0.6	75.8 $\pm$ 0.5	75.8 $\pm$ 0.75.0 $\pm$ 0.7	75.8 $\pm$ 0.5	43.6 $\pm$ 0.40.9 $\pm$ 0.4	82.6 $\pm$ 0.5	53.8 $\pm$ 0.45.1 $\pm$ 0.4	91.2 $\pm$ 0.5
	GCN-HEAT	90.5 $\pm$ 0.48.5 $\pm$ 0.5	90.3 $\pm$ 0.5	84.5 $\pm$ 0.48.0 $\pm$ 0.5	87.4 $\pm$ 0.5	48.2 $\pm$ 0.47.0 $\pm$ 0.3	75.8 $\pm$ 0.5	54.5 $\pm$ 0.64.4 $\pm$ 0.4	75.8 $\pm$ 0.5	75.8 $\pm$ 0.82.9 $\pm$ 0.9	75.8 $\pm$ 0.5	43.6 $\pm$ 0.43.8 $\pm$ 0.4	82.6 $\pm$ 0.5	53.8 $\pm$ 0.42.0 $\pm$ 0.4	91.2 $\pm$ 0.5
	GCNsafe	90.5 $\pm$ 0.49.1 $\pm$ 0.5	90.3 $\pm$ 0.5	84.5 $\pm$ 0.48.1 $\pm$ 0.5	87.4 $\pm$ 0.5	48.2 $\pm$ 0.47.0 $\pm$ 0.3	75.8 $\pm$ 0.5	54.5 $\pm$ 0.64.8 $\pm$ 0.4	75.8 $\pm$ 0.5	75.8 $\pm$ 0.75.0 $\pm$ 0.7	75.8 $\pm$ 0.5	43.6 $\pm$ 0.43.9 $\pm$ 0.4	82.6 $\pm$ 0.5	53.8 $\pm$ 0.45.1 $\pm$ 0.4	91.2 $\pm$ 0.5
	GCN-GBM	90.5 $\pm$ 0.49.1 $\pm$ 0.5	90.3 $\pm$ 0.5	84.5 $\pm$ 0.48.2 $\pm$ 0.5	87.4 $\pm$ 0.5	48.2 $\pm$ 0.47.1 $\pm$ 0.3	75.8 $\pm$ 0.5	54.5 $\pm$ 0.64.8 $\pm$ 0.4	75.8 $\pm$ 0.5	75.8 $\pm$ 0.82.9 $\pm$ 0.9	75.8 $\pm$ 0.5	43.6 $\pm$ 0.43.9 $\pm$ 0.4	82.6 $\pm$ 0.5	53.8 $\pm$ 0.45.1 $\pm$ 0.4	91.2 $\pm$ 0.5
CoraML-LLM	GCN-MCD	89.5 $\pm$ 0.45.1 $\pm$ 0.5	90.1 $\pm$ 0.5	84.1 $\pm$ 0.439.1 $\pm$ 0.4	89.5 $\pm$ 0.5	n.a./n.a.	n.a.	48.3 $\pm$ 0.25.5 $\pm$ 0.3	76.9 $\pm$ 0.5	65.2 $\pm$ 0.464.1 $\pm$ 0.4	76.9 $\pm$ 0.5	51.5 $\pm$ 0.21.4 $\pm$ 0.2	84.0 $\pm$ 0.5	54.5 $\pm$ 0.22.0 $\pm$ 0.2	91.0 $\pm$ 0.5
	GCN-DE	88.1 $\pm$ 0.47.0 $\pm$ 0.5	87.8 $\pm$ 0.5	82.4 $\pm$ 0.460.1 $\pm$ 0.4	87.9 $\pm$ 0.5	n.a./n.a.	n.a.	54.2 $\pm$ 0.31.9 $\pm$ 0.3	75.6 $\pm$ 0.5	66.3 $\pm$ 0.477.2 $\pm$ 0.4	75.6 $\pm$ 0.5	54.2 $\pm$ 0.39.5 $\pm$ 0.3	81.8 $\pm$ 0.5	52.6 $\pm$ 0.36.6 $\pm$ 0.3	88.5 $\pm$ 0.5
	GCN-MCD+DE	87.6 $\pm$ 0.47.1 $\pm$ 0.5	88.1 $\pm$ 0.5	81.7 $\pm$ 0.460.3 $\pm$ 0.3	87.9 $\pm$ 0.5	n.a./n.a.	n.a.	52.9 $\pm$ 0.42.3 $\pm$ 0.4	75.2 $\pm$ 0.5	66.3 $\pm$ 0.478.3 $\pm$ 0.4	75.2 $\pm$ 0.5	51.5 $\pm$ 0.27.6 $\pm$ 0.2	81.3 $\pm$ 0.5	52.4 $\pm$ 0.37.3 $\pm$ 0.3	88.7 $\pm$ 0.5
	GCN-BNN	89.5 $\pm$ 0.479.3 $\pm$ 0.4	90.0 $\pm$ 0.5	82.9 $\pm$ 0.475.8 $\pm$ 0.6	89.3 $\pm$ 0.5	n.a./n.a.	n.a.	46.7 $\pm$ 0.26.4 $\pm$ 0.3	76.8 $\pm$ 0.5	65.8 $\pm$ 0.474.9 $\pm$ 0.4	76.8 $\pm$ 0.5	51.6 $\pm$ 0.43.8 $\pm$ 0.4	83.7 $\pm$ 0.5	58.5 $\pm$ 0.49.0 $\pm$ 0.4	90.8 $\pm$ 0.5
	GCN-Ess	89.5 $\pm$ 0.477.5 $\pm$ 0.4	90.1 $\pm$ 0.5	84.1 $\pm$ 0.470.0 $\pm$ 0.4	89.5 $\pm$ 0.5	n.a./n.a.	n.a.	46.6 $\pm$ 0.21.4 $\pm$ 0.2	77.3 $\pm$ 0.5	63.8 $\pm$ 0.474.8 $\pm$ 0.4	77.3 $\pm$ 0.5	51.6 $\pm$ 0.43.1 $\pm$ 0.3	84.2 $\pm$ 0.5	53.8 $\pm$ 0.37.3 $\pm$ 0.3	91.3 $\pm$ 0.5
	GCN	85.0 $\pm$ 0.488.9 $\pm$ 0.4	89.4 $\pm$ 0.5	83.5 $\pm$ 0.477.5 $\pm$ 0.4	88.1 $\pm$ 0.5	n.a./n.a.	n.a.	53.1 $\pm$ 0.48.8 $\pm$ 0.5	75.8 $\pm$ 0.5	55.3 $\pm$ 0.41.1 $\pm$ 0.1	75.8 $\pm$ 0.5	54.4 $\pm$ 0.36.6 $\pm$ 0.3	81.3 $\pm$ 0.5	61.6 $\pm$ 0.44.8 $\pm$ 0.4	90.8 $\pm$ 0.5
	SGCN	90.5 $\pm$ 0.486.7 $\pm$ 0.5	88.9 $\pm$ 0.5	83.1 $\pm$ 0.477.1 $\pm$ 0.7	89.3 $\pm$ 0.5	n.a./n.a.	n.a.	59.7 $\pm$ 0.48.0 $\pm$ 0.5	68.1 $\pm$ 0.5	69.1 $\pm$ 0.466.7 $\pm$ 0.6	68.1 $\pm$ 0.5	54.6 $\pm$ 0.36.6 $\pm$ 0.3	78.9 $\pm$ 0.5	60.0 $\pm$ 0.46.3 $\pm$ 0.3	80.2 $\pm$ 0.5
	GCN-EBM	89.4 $\pm$ 0.486.7 $\pm$ 0.5	90.2 $\pm$ 0.5	83.8 $\pm$ 0.483.6 $\pm$ 0.4	89.4 $\pm$ 0.5	n.a./n.a.	n.a.	47.5 $\pm$ 0.42.9 $\pm$ 0.4	77.1 $\pm$ 0.5	67.2 $\pm$ 0.465.9 $\pm$ 0.6	77.1 $\pm$ 0.5	51.9 $\pm$ 0.44.5 $\pm$ 0.4	83.7 $\pm$ 0.5	54.6 $\pm$ 0.36.6 $\pm$ 0.3	91.1 $\pm$ 0.5
	GCN-HEAT	89.4 $\pm$ 0.486.7 $\pm$ 0.5	90.2 $\pm$ 0.5	83.8 $\pm$ 0.477.2 $\pm$ 0.7	89.4 $\pm$ 0.5	n.a./n.a.	n.a.	47.5 $\pm$ 0.46.2 $\pm$ 0.6	77.1 $\pm$ 0.5	67.2 $\pm$ 0.476.2 $\pm$ 0.6	77.1 $\pm$ 0.5	51.9 $\pm$ 0.431.9 $\pm$ 0.2	83.7 $\pm$ 0.5	54.6 $\pm$ 0.39.3 $\pm$ 0.3	91.1 $\pm$ 0.5
	GCNsafe	89.4 $\pm$ 0.486.7 $\pm$ 0.5	90.2 $\pm$ 0.5	83.8 $\pm$ 0.486.6 $\pm$ 0.4	89.4 $\pm$ 0.5	n.a./n.a.	n.a.	47.5 $\pm$ 0.46.1 $\pm$ 0.7	77.1 $\pm$ 0.5	67.2 $\pm$ 0.465.3 $\pm$ 0.6	77.1 $\pm$ 0.5	51.9 $\pm$ 0.43.8 $\pm$ 0.4	83.7 $\pm$ 0.5	54.6 $\pm$ 0.32.0 $\pm$ 0.2	91.1 $\pm$ 0.5
	GCN-GBM	89.4 $\pm$ 0.486.7 $\pm$ 0.5	90.2 $\pm$ 0.5	83.8 $\pm$ 0.486.6 $\pm$ 0.4	89.4 $\pm$ 0.5	n.a./n.a.	n.a.	47.5 $\pm$ 0.46.1 $\pm$ 0.7	77.1 $\pm$ 0.5	67.2 $\pm$ 0.465.3 $\pm$ 0.6	77.1 $\pm$ 0.5	51.9 $\pm$ 0.43.8 $\pm$ 0.4	83.7 $\pm$ 0.5	54.6 $\pm$ 0.32.0 $\pm$ 0.2	91.1 $\pm$ 0.5
Cluster	GCN-MCD	84.2 $\pm$ 0.417.7 $\pm$ 0.5	84.8 $\pm$ 0.5	86.2 $\pm$ 0.441.1 $\pm$ 0.4	90.9 $\pm$ 0.5	54.2 $\pm$ 0.436.3 $\pm$ 0.3	76.4 $\pm$ 0.5	63.4 $\pm$ 0.44.9 $\pm$ 0.4	76.4 $\pm$ 0.5	67.8 $\pm$ 0.471.3 $\pm$ 0.3	76.4 $\pm$ 0.5	50.2 $\pm$ 0.21.0 $\pm$ 0.2	78.4 $\pm$ 0.5	46.4 $\pm$ 0.20.2 $\pm$ 0.2	77.8 $\pm$ 0.5
	GCN-DE	78.8 $\pm$ 0.465.7 $\pm$ 0.5	81.0 $\pm$ 0.5	83.3 $\pm$ 0.466.9 $\pm$ 0.3	86.4 $\pm$ 0.5	56.3 $\pm$ 0.459.5 $\pm$ 0.5	72.9 $\pm$ 0.5	55.8 $\pm$ 0.473.9 $\pm$ 0.5	72.9 $\pm$ 0.5	63.9 $\pm$ 0.478.2 $\pm$ 0.4	72.9 $\pm$ 0.5	49.6 $\pm$ 0.42.8 $\pm$ 0.4	75.6 $\pm$ 0.5	48.8 $\pm$ 0.41.6 $\pm$ 0.3	74.3 $\pm$ 0.5
	GCN-MCD+DE	79.3 $\pm$ 0.466.5 $\pm$ 0.5	81.3 $\pm$ 0.5	83.9 $\pm$ 0.467.4 $\pm$ 0.4	87.1 $\pm$ 0.5	54.6 $\pm$ 0.459.4 $\pm$ 0.5	73.0 $\pm$ 0.5	55.2 $\pm$ 0.472.3 $\pm$ 0.3	73.0 $\pm$ 0.5	64.7 $\pm$ 0.479.0 $\pm$ 0.3	73.0 $\pm$ 0.5	49.4 $\pm$ 0.41.8 $\pm$ 0.4	75.0 $\pm$ 0.5	49.3 $\pm$ 0.42.2 $\pm$ 0.2	74.6 $\pm$ 0.5
	GCN-BNN	84.1 $\pm$ 0.471.5 $\pm$ 0.4	84.8 $\pm$ 0.5	85.9 $\pm$ 0.472.8 $\pm$ 0.3	91.1 $\pm$ 0.5	57.9 $\pm$ 0.439.2 $\pm$ 0.4	76.9 $\pm$ 0.5	59.1 $\pm$ 0.462.6 $\pm$ 0.6	76.9 $\pm$ 0.5	68.0 $\pm$ 0.477.8 $\pm$ 0.4	76.9 $\pm$ 0.5	49.2 $\pm$ 0.41.7 $\pm$ 0.3	78.1 $\pm$ 0.5	46.2 $\pm$ 0.33.6 $\pm$ 0.3	74.1 $\pm$ 0.5
	GCN-Ess	84.1 $\pm$ 0.471.5 $\pm$ 0.4	84.8 $\pm$ 0.5	85.9 $\pm$ 0.472.8 $\pm$ 0.3	91.1 $\pm$ 0.5	57.9 $\pm$ 0.439.2 $\pm$ 0.4	76.9 $\pm$ 0.5	59.1 $\pm$ 0.462.6 $\pm$ 0.6	76.9 $\pm$ 0.5	68.0 $\pm$ 0.477.8 $\pm$ 0.4	76.9 $\pm$ 0.5	49.2 $\pm$ 0.41.7 $\pm$ 0.3	78.1 $\pm$ 0.5	46.2 $\pm$ 0.33.6 $\pm$ 0.3	74.1 $\pm$ 0.5
	GCN	84.7 $\pm$ 0.484.0 $\pm$ 0.5	85.6 $\pm$ 0.5	88.8 $\pm$ 0.480.0 $\pm$ 0.4	91.6 $\pm$ 0.5	57.8 $\pm$ 0.453.0 $\pm$ 0.4	76.7 $\pm$ 0.5	61.7 $\pm$ 0.464.7 $\pm$ 0.4	76.7 $\pm$ 0.5	64.5 $\pm$ 0.466.0 $\pm$ 0.3	76.7 $\pm$ 0.5	49.2 $\pm$ 0.41.7 $\pm$ 0.3	78.1 $\pm$ 0.5	46.2 $\pm$ 0.33.6 $\pm$ 0.3	74.1 $\pm$ 0.5
	SGCN	83.8 $\pm$ 0.479.7 $\pm$ 0.5	85.0 $\pm$ 0.5	84.5 $\pm$ 0.478.7 $\pm$ 0.7	90.4 $\pm$ 0.5	58.9 $\pm$ 0.454.0 $\pm$ 0.4	75.8 $\pm$ 0.5	63.4 $\pm$ 0.469.8 $\pm$ 0.6	75.8 $\pm$ 0.5	69.6 $\pm$ 0.468.5 $\pm$ 0.6	75.8 $\pm$ 0.5	49.0 $\pm$ 0.48.4 $\pm$ 0.4	77.1 $\pm$ 0.5	51.0 $\pm$ 0.47.7 $\pm$ 0.7	76.6 $\pm$ 0.5
	GCN-EBM	84.3 $\pm$ 0.484.1 $\pm$ 0.5	84.8 $\pm$ 0.5	86.5 $\pm$ 0.485.5 $\pm$ 0.5	90.9 $\pm$ 0.5	56.7 $\pm$ 0.456.9 $\pm$ 0.6	76.6 $\pm$ 0.5	61.2 $\pm$ 0.464.4 $\pm$ 0.4	76.6 $\pm$ 0.5	68.8 $\pm$ 0.466.8 $\pm$ 0.6	76.6 $\pm$ 0.5	49.5 $\pm$ 0.48.6 $\pm$ 0.6	78.7 $\pm$ 0.5	46.7 $\pm$ 0.47.8 $\pm$ 0.7	77.5 $\pm$ 0.5
	GCN-HEAT	84.3 $\pm$ 0.477.9 $\pm$ 0.5	84.8 $\pm$ 0.5	86.5 $\pm$ 0.479.8 $\pm$ 0.3	90.9 $\pm$ 0.5	56.6 $\pm$ 0.46.7 $\pm$ 0.6	76.6 $\pm$ 0.5	62.1 $\pm$ 0.460.7 $\pm$ 0.6	76.6 $\pm$ 0.5	68.8 $\pm$ 0.480.3 $\pm$ 0.6	76.6 $\pm$ 0.5	49.5 $\pm$ 0.49.7 $\pm$ 0.4	78.7 $\pm$ 0.5	46.7 $\pm$	

	Model	LoC (hans)		LoC (hetem)		AUC(Ber)		AUC(Ber0.5)		AUC(0.1) (fur)		Page Rank		Homophily	
		AUC-PR†	Acc.†	AUC-PR†	Acc.†	AUC-PR†	Acc.†	AUC-PR†	Acc.†	AUC-PR†	Acc.†	AUC-PR†	Acc.†	AUC-PR†	Acc.†
		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)		(Alea. / Epi.)	
CoMIL	GCN-MCD	88.61±0.51,59.11	90.41±0.8	80.41±0.43,3.27	85.21±0.3	53.51±0.29,6.12	82.01±0.4	54.81±0.32,0.20	82.01±0.4	76.61±0.77,6.12	82.01±0.4	45.31±0.22,6.12	82.61±0.4	52.81±0.22,2.14	93.91±0.4
	GCN-DE	86.31±0.71,61.21	87.31±0.2	78.71±0.59,9.26	80.31±0.2	51.51±0.45,4.71	78.41±0.9	52.61±0.58,0.21	78.41±0.9	72.81±0.81,0.21	78.41±0.9	46.51±0.31,1.21	79.01±0.3	50.61±0.36,4.19	90.91±0.7
	GCN-MCD+DE	86.71±0.72,81.11	87.61±0.2	78.31±0.61,2.20	80.11±0.3	51.81±0.55,3.19	78.11±0.2	53.21±0.59,7.18	78.11±0.2	70.41±0.80,7.18	78.11±0.2	46.51±0.37,4.19	79.51±0.7	51.81±0.37,0.11	90.81±0.7
	GCN-BNN	88.21±0.86,11.11	89.01±0.2	76.91±0.73,7.11	82.11±0.4	55.71±0.50,8.18	80.61±0.2	64.51±0.64,7.18	80.61±0.2	81.01±0.83,9.17	80.61±0.2	48.61±0.44,2.19	80.41±0.2	53.41±0.50,1.18	93.11±0.4
	GCN-Ems	88.81±0.74,8.11	90.31±0.0	80.91±0.53,3.12	85.21±0.3	53.11±0.29,3.12	81.81±0.9	54.21±0.34,1.11	81.81±0.9	76.71±0.81,6.12	81.81±0.9	44.61±0.27,8.17	82.71±0.3	53.41±0.33,3.11	94.01±0.3
	GCN	86.11±0.87,7.11	90.01±0.0	77.71±0.77,7.11	85.51±0.3	57.11±0.41,8.17	78.81±0.9	56.31±0.40,9.11	78.81±0.9	58.01±0.40,4.11	78.81±0.9	45.51±0.30,5.11	78.41±0.9	55.61±0.41,9.11	91.11±0.3
	SGCN	88.71±0.89,4.11	91.01±0.2	82.01±0.83,2.11	88.51±0.2	75.31±0.63,2.17	51.71±0.3	73.71±0.70,5.17	51.71±0.3	79.81±0.79,1.17	51.71±0.3	49.01±0.38,1.11	83.31±0.3	53.31±0.43,3.11	94.21±0.2
	GCN-EBM	88.71±0.88,2.11	90.41±0.0	80.71±0.80,1.11	85.01±0.3	50.81±0.48,5.11	81.91±0.4	54.31±0.52,3.11	81.91±0.4	77.21±0.75,6.11	81.91±0.4	45.31±0.41,7.11	82.51±0.4	52.81±0.50,4.11	94.11±0.0
	GCN-HEAT	88.71±0.85,2.11	90.41±0.0	80.71±0.74,6.01	85.01±0.3	50.81±0.46,1.17	81.91±0.4	54.31±0.45,3.11	81.91±0.4	77.21±0.83,3.11	81.91±0.4	45.31±0.39,7.11	82.51±0.4	52.81±0.47,0.11	94.11±0.0
	GCN-Safe	88.71±0.90,2.11	90.41±0.0	80.71±0.82,3.11	85.01±0.3	50.81±0.51,5.11	81.91±0.4	54.31±0.53,3.11	81.91±0.4	77.21±0.75,6.11	81.91±0.4	45.31±0.40,1.11	82.51±0.4	52.81±0.47,0.11	94.11±0.0
GCN-GERM	88.71±0.89,3.11	90.41±0.0	80.71±0.82,3.11	85.01±0.3	50.81±0.46,8.11	81.91±0.4	54.31±0.45,7.11	81.91±0.4	77.21±0.81,4.11	81.91±0.4	45.31±0.39,2.11	82.51±0.4	52.81±0.43,9.11	94.11±0.0	
CoMIL+LM	GCN-MCD	89.31±0.70,41.11	91.21±0.7	81.41±0.89,4.11	90.61±0.2	n.a./n.a.	n.a.	48.71±0.24,5.11	82.61±0.3	65.21±0.63,4.11	82.61±0.3	50.11±0.19,9.11	84.91±0.3	50.71±0.19,9.11	94.11±0.3
	GCN-DE	89.71±0.73,9.11	89.71±0.2	80.61±0.62,2.11	88.91±0.7	n.a./n.a.	n.a.	53.11±0.33,3.11	79.91±0.4	66.71±0.77,4.11	79.91±0.4	51.81±0.37,1.11	82.21±0.7	55.71±0.39,8.11	91.71±0.3
	GCN-MCD+DE	89.21±0.73,9.11	89.41±0.0	80.81±0.61,4.11	88.91±0.7	n.a./n.a.	n.a.	53.01±0.35,3.11	80.31±0.4	66.01±0.76,2.11	80.31±0.4	51.51±0.38,2.11	82.71±0.3	58.31±0.40,1.11	91.41±0.3
	GCN-BNN	89.21±0.82,7.11	91.41±0.0	80.11±0.73,6.01	91.01±0.2	n.a./n.a.	n.a.	47.81±0.28,9.11	82.81±0.4	68.11±0.73,7.11	82.81±0.4	51.61±0.42,3.11	84.01±0.2	61.11±0.53,7.11	94.01±0.3
	GCN-Ems	89.41±0.83,8.11	91.41±0.0	81.61±0.67,0.11	90.61±0.2	n.a./n.a.	n.a.	49.11±0.27,5.11	82.61±0.3	66.01±0.76,4.11	82.61±0.3	49.51±0.32,8.11	85.01±0.3	60.31±0.44,3.11	94.01±0.3
	GCN	87.41±0.80,3.11	90.51±0.2	82.21±0.76,5.11	88.51±0.2	n.a./n.a.	n.a.	53.71±0.42,6.11	79.71±0.4	53.21±0.42,0.11	79.71±0.4	54.61±0.49,1.11	80.91±0.2	58.11±0.44,3.11	92.51±0.3
	SGCN	90.91±0.90,4.11	90.31±0.0	83.11±0.83,4.11	90.61±0.2	n.a./n.a.	n.a.	65.11±0.43,5.11	68.71±0.2	69.41±0.67,0.11	82.01±0.4	54.41±0.39,3.11	82.01±0.4	56.71±0.46,4.11	93.11±0.3
	GCN-EBM	89.21±0.88,3.11	91.41±0.0	81.51±0.81,4.11	90.71±0.2	n.a./n.a.	n.a.	47.61±0.43,7.11	82.61±0.3	67.11±0.64,6.11	82.61±0.3	50.41±0.44,9.11	85.01±0.3	60.91±0.57,4.11	93.81±0.3
	GCN-HEAT	89.21±0.82,5.11	91.41±0.0	81.51±0.77,4.11	90.71±0.2	n.a./n.a.	n.a.	47.61±0.38,4.11	82.61±0.3	67.11±0.72,8.11	82.61±0.3	50.41±0.38,6.11	85.01±0.3	60.91±0.46,3.11	93.81±0.3
	GCN-Safe	89.21±0.90,6.11	91.41±0.0	81.51±0.86,4.11	90.71±0.2	n.a./n.a.	n.a.	47.61±0.50,2.11	82.61±0.3	67.11±0.65,2.11	82.61±0.3	50.41±0.45,3.11	85.01±0.3	60.91±0.52,2.11	93.81±0.3
GCN-GERM	89.21±0.89,6.11	91.41±0.0	81.51±0.81,6.11	90.71±0.2	n.a./n.a.	n.a.	47.61±0.37,4.11	82.61±0.3	67.11±0.70,6.11	82.61±0.3	50.41±0.40,5.11	85.01±0.3	60.91±0.50,9.11	93.81±0.3	
Chisnet	GCN-MCD	81.81±0.41,9.11	83.51±0.3	82.81±0.42,7.11	91.11±0.8	56.21±0.36,7.11	82.11±0.4	61.11±0.39,0.11	82.11±0.4	69.01±0.77,0.11	82.11±0.4	47.01±0.19,5.11	82.91±0.3	50.21±0.19,2.11	86.31±0.3
	GCN-DE	79.11±0.67,7.11	80.11±0.2	81.11±0.66,7.11	96.41±0.1	56.21±0.36,7.11	80.71±0.2	57.81±0.31,8.11	80.71±0.2	65.31±0.81,8.11	80.71±0.2	49.31±0.45,9.11	79.01±0.3	49.31±0.45,9.11	84.81±0.3
	GCN-MCD+DE	78.81±0.66,1.11	80.21±0.2	81.21±0.68,3.11	86.21±0.1	55.81±0.46,7.11	76.81±0.3	58.51±0.43,3.11	76.81±0.3	65.61±0.78,8.11	76.81±0.3	48.11±0.42,5.11	79.21±0.3	50.71±0.45,9.11	81.71±0.3
	GCN-BNN	89.21±0.79,3.11	82.51±0.4	81.41±0.80,4.11	90.21±0.7	55.31±0.62,2.11	81.21±0.3	63.41±0.75,5.11	81.21±0.3	71.91±0.79,7.11	81.21±0.3	46.81±0.47,8.11	81.31±0.3	50.01±0.50,3.11	86.01±0.3
	GCN-Ems	81.61±0.70,1.11	83.81±0.3	82.61±0.73,4.11	91.21±0.8	54.01±0.49,3.11	82.01±0.4	60.11±0.59,8.11	82.01±0.4	69.21±0.77,5.11	82.01±0.4	47.11±0.35,3.11	82.61±0.3	49.21±0.49,4.11	86.41±0.3
	GCN	81.51±0.80,5.11	84.91±0.2	86.01±0.78,9.11	90.91±0.8	60.11±0.57,9.11	83.31±0.3	60.11±0.59,8.11	83.31±0.3	67.61±0.82,3.11	83.31±0.3	54.51±0.44,7.11	82.11±0.3	55.01±0.47,2.11	88.01±0.3
	SGCN	88.21±0.85,5.11	83.51±0.3	84.81±0.86,6.11	90.61±0.2	60.41±0.54,8.11	81.51±0.3	65.91±0.61,4.11	81.51±0.3	69.21±0.68,3.11	81.51±0.3	48.51±0.43,3.11	82.51±0.3	51.01±0.48,7.11	85.91±0.3
	GCN-EBM	81.61±0.81,2.11	83.61±0.3	82.31±0.81,3.11	91.11±0.8	55.61±0.54,4.11	82.01±0.4	60.01±0.53,6.11	82.01±0.4	70.11±0.66,9.11	82.01±0.4	47.31±0.45,1.11	82.41±0.3	49.21±0.46,9.11	86.41±0.3
	GCN-HEAT	81.61±0.74,6.11	83.61±0.3	82.31±0.75,2.11	91.11±0.8	55.61±0.54,4.11	82.01±0.4	60.01±0.53,6.11	82.01±0.4	70.11±0.66,9.11	82.01±0.4	47.31±0.45,1.11	82.41±0.3	49.21±0.46,9.11	86.41±0.3
	GCN-Safe	81.61±0.84,2.11	83.61±0.3	82.31±0.81,9.11	91.11±0.8	55.61±0.63,1.11	82.01±0.4	60.01±0.62,8.11	82.01±0.4	70.11±0.66,2.11	82.01±0.4	47.31±0.49,9.11	82.41±0.3	49.21±0.46,2.11	86.41±0.3
GCN-GERM	81.61±0.82,8.11	83.61±0.3	82.31±0.82,3.11	91.11±0.8	55.61±0.57,2.11	82.01±0.4	60.01±0.62,7.11	82.01±0.4	70.11±0.66,9.11	82.01±0.4	47.31±0.46,0.11	82.41±0.3	49.21±0.46,9.11	86.41±0.3	
Amazon Computers	GCN-MCD	66.31±0.43,8.11	91.61±0.6	53.41±0.32,3.11	83.71±0.3	n.a./n.a.	n.a.	47.61±0.39,5.11	78.11±0.3	57.41±0.56,0.11	78.11±0.3	41.31±0.21,8.11	77.91±0.3	42.31±0.20,2.11	88.91±0.3
	GCN-DE	65.91±0.55,2.11	91.21±0.7	54.81±0.47,5.11	80.51±0.3	n.a./n.a.	n.a.	47.61±0.31,6.11	71.91±0.3	55.51±0.60,2.11	71.91±0.3	40.51±0.32,5.11	76.01±0.3	42.61±0.26,2.11	84.81±0.3
	GCN-MCD+DE	66.61±0.55,4.11	90.81±0.3	55.51±0.47,9.11	80.71±0.3	n.a./n.a.	n.a.	47.61±0.31,6.11	75.41±0.3	55.01±0.59,4.11	75.41±0.3	40.11±0.33,7.11	76.31±0.3	42.91±0.26,0.11	84.91±0.3
	GCN-BNN	66.31±0.64,6.11	90.71±0.6	53.41±0.42,7.11	84.01±0.7	n.a./n.a.	n.a.	51.21±0.47,5.11	78.21±0.3	59.11±0.61,5.11	78.21±0.3	43.51±0.40,9.11	76.11±0.3	42.61±0.41,8.11	89.31±0.3
	GCN-Ems	66.21±0.50,9.11	91.71±0.3	53.31±0.44,6.11	83.81±0.4	n.a./n.a.	n.a.	47.51±0.36,7.11	78.11±0.3	58.41±0.61,4.11	78.11±0.3	41.41±0.23,9.11	77.91±0.3	42.11±0.30,3.11	88.91±0.3
	GCN	64.91±0.62,6.11	92.31±0.0	53.11±0.55,1.11	85.01±0.3	n.a./n.a.	n.a.	48.31±0.38,6.11	79.81±0.3	43.61±0.39,6.11	79.81±0.3	43.21±0.51,5.11	80.21±0.3	43.61±0.32,9.11	90.31±0.3
	SGCN	66.31±0.68,6.11	92.11±0.6	54.31±0.56,4.11	85.11±0.0	n.a./n.a.	n.a.	48.11±0.39,0.11	80.11±0.3	57.21±0.56,1.11	80.11±0.3	42.01±0.34,4.11	79.31±0.3	41.31±0.36,0.11	90.71±0.3
	GCN-EBM	66.21±0.66,3.11	91.71±0.3	53.41±0.52,9.11	83.71±0.3	n.a./n.a.	n.a.	48.41±0.43,2.11	78.11±0.3	58.51±0.58,3.11	78.11±0.3	41.21±0.28,4.11	78.01±0.3	42.61±0.40,2.11	88.81±0.3
	GCN-HEAT	66.21±0.69,9.11	91.71±0.3	53.41±0.46,6.11	83.71±0.3	n.a./n.a.	n.a.	48.61±0.46,0.11	78.11±0.3	58.51±0.62,0.11	78.11±0.3	41.21±0.29,9.11	78.01±0.3	42.61±0.43,3.11	88.81±0.3
	GCN-Safe	66.21±0.65,3.11	91.71±0.3	53.41±0.56,7.11	83.71±0.3	n.a./n.a.	n.a.	48.61±0.46,1.11	78.11±0.3	58.51±0.57,8.11	78.11±0.3	41.21±0.37,2.11	78.01±0.3	42.61±0.37,7.11	88.81±0.3
GCN-GERM	66.21±0.63,2.11	91.71±0.3	53.41±0.51,6.11	83.71±0.3	n.a./n.a.	n.a.	48.61±0.39,4.11	78.11±0.3	58.51±0.61,1.11	78.11±0.3	41.21±0.28,4.11	78.01±0.3	42.61±0.32,8.11	88.81±0.3	
Amazon Photo	GCN-MCD	74.31±0.46,11.11	92.01±0.6	66.91±0.41,1.11	96.31±0.1	38.11±0.17,6.11	90.21±0.6	64.71±0.47,0.11	90.21±0.6	65.91±0.41,1.11	90.21±0.6	11.01±0.14,7.11	91.81±0.1	46.11±0.12,8.11	97.61±0.1
	GCN-DE	75.71±0.57,12.11	91.01±0.6	66.21±0.50,6.11	95.91±0.6	38.11±0.17,9.11	89.51±0.6	64.11±0.46,7.11	89.51±0.6	69.91±0.46,1.11	89.51±0.6	10.51±0.14,7.11	91.21±0.1	41.11±0.28,8.11	96.91±0

	Model	LoC (i.i.d.) Clean / I.d. / O.o.D.	LoC (heterophilic) Clean / I.d. / O.o.D.	Bert(β) (near) Clean / I.d. / O.o.D.	Bert(0.5) Clean / I.d. / O.o.D.	$\mathcal{N}(0, 1)$ (far) Clean / I.d. / O.o.D.	Page Rank Clean / I.d. / O.o.D.	Homophily Clean / I.d. / O.o.D.
CoMML	BGCN	6.4 _{1.1} / 5.7 _{1.1} / 60.7 _{1.1}	7.5 _{1.1} / 7.1 _{1.1} / 68.0 _{1.1}	12.2 _{1.1} / 10.6 _{1.1} / 12.1 _{1.1}	12.2 _{1.1} / 10.6 _{1.1} / 19.3 _{1.1}	12.2 _{1.1} / 48.5 _{1.1} / 56.7 _{1.1}	11.8 _{1.1} / 11.2 _{1.1} / 8.6 _{1.1}	4.9 _{1.1} / 5.8 _{1.1} / 9.4 _{1.1}
	GCN-MCD	6.1 _{1.1} / 7.2 _{1.1} / 51.3 _{1.1}	7.0 _{1.1} / 8.7 _{1.1} / 49.8 _{1.1}	7.7 _{1.1} / 27.1 _{1.1} / 30.3 _{1.1}	7.7 _{1.1} / 23.1 _{1.1} / 21.2 _{1.1}	7.7 _{1.1} / 42.8 _{1.1} / 65.3 _{1.1}	6.8 _{1.1} / 15.5 _{1.1} / 15.1 _{1.1}	12.0 _{1.1} / 21.0 _{1.1} / 15.1 _{1.1}
	GCN-DE	11.3 _{1.1} / 13.0 _{1.1} / 47.2 _{1.1}	11.7 _{1.1} / 12.5 _{1.1} / 46.4 _{1.1}	16.7 _{1.1} / 28.4 _{1.1} / 24.0 _{1.1}	16.7 _{1.1} / 21.8 _{1.1} / 12.9 _{1.1}	16.7 _{1.1} / 34.0 _{1.1} / 57.3 _{1.1}	12.4 _{1.1} / 12.0 _{1.1} / 19.4 _{1.1}	22.3 _{1.1} / 29.9 _{1.1} / 21.1 _{1.1}
	GCN-Ens	6.0 _{1.1} / 7.6 _{1.1} / 51.3 _{1.1}	7.0 _{1.1} / 8.3 _{1.1} / 50.0 _{1.1}	7.2 _{1.1} / 26.3 _{1.1} / 31.7 _{1.1}	7.2 _{1.1} / 24.3 _{1.1} / 21.5 _{1.1}	7.2 _{1.1} / 40.4 _{1.1} / 59.4 _{1.1}	6.5 _{1.1} / 6.9 _{1.1} / 15.9 _{1.1}	12.4 _{1.1} / 21.6 _{1.1} / 15.1 _{1.1}
	GCN-MCD+DE	11.4 _{1.1} / 12.8 _{1.1} / 47.2 _{1.1}	12.3 _{1.1} / 13.3 _{1.1} / 46.5 _{1.1}	17.1 _{1.1} / 29.4 _{1.1} / 23.7 _{1.1}	17.1 _{1.1} / 22.5 _{1.1} / 13.4 _{1.1}	17.1 _{1.1} / 37.8 _{1.1} / 56.5 _{1.1}	12.1 _{1.1} / 12.3 _{1.1} / 30.2 _{1.1}	22.3 _{1.1} / 29.9 _{1.1} / 21.7 _{1.1}
	SGCN	32.9 _{1.1} / 24.9 _{1.1} / 34.0 _{1.1}	31.5 _{1.1} / 33.2 _{1.1} / 38.9 _{1.1}	18.8 _{1.1} / 18.1 _{1.1} / 15.9 _{1.1}	18.8 _{1.1} / 16.1 _{1.1} / 13.4 _{1.1}	18.8 _{1.1} / 64.4 _{1.1} / 105.1 _{1.1}	39.0 _{1.1} / 28.3 _{1.1} / 44.8 _{1.1}	54.7 _{1.1} / 60.0 _{1.1} / 46.2 _{1.1}
	SPN	6.0 _{1.1} / 7.4 _{1.1} / 54.2 _{1.1}	9.6 _{1.1} / 10.9 _{1.1} / 52.0 _{1.1}	9.7 _{1.1} / 15.1 _{1.1} / 17.5 _{1.1}	9.7 _{1.1} / 14.9 _{1.1} / 15.9 _{1.1}	9.7 _{1.1} / 12.1 _{1.1} / 13.2 _{1.1}	12.9 _{1.1} / 12.7 _{1.1} / 13.0 _{1.1}	5.1 _{1.1} / 9.6 _{1.1} / 15.9 _{1.1}
	GCN (Ours)	6.2 _{1.1} / 7.8 _{1.1} / 51.4 _{1.1}	6.9 _{1.1} / 8.3 _{1.1} / 50.2 _{1.1}	7.8 _{1.1} / 26.4 _{1.1} / 31.5 _{1.1}	7.8 _{1.1} / 24.4 _{1.1} / 21.1 _{1.1}	7.8 _{1.1} / 45.0 _{1.1} / 64.8 _{1.1}	7.0 _{1.1} / 6.9 _{1.1} / 16.3 _{1.1}	12.1 _{1.1} / 21.2 _{1.1} / 14.7 _{1.1}
	BGCN	4.8 _{1.1} / 4.8 _{1.1} / 69.3 _{1.1}	4.6 _{1.1} / 4.8 _{1.1} / 64.2 _{1.1}	n.a. / n.a. / n.a.	7.4 _{1.1} / 19.1 _{1.1} / 30.9 _{1.1}	7.4 _{1.1} / 16.4 _{1.1} / 30.9 _{1.1}	5.8 _{1.1} / 6.5 _{1.1} / 7.6 _{1.1}	4.6 _{1.1} / 5.8 _{1.1} / 13.4 _{1.1}
	GCN-MCD	4.0 _{1.1} / 4.2 _{1.1} / 62.0 _{1.1}	4.1 _{1.1} / 5.3 _{1.1} / 60.3 _{1.1}	n.a. / n.a. / n.a.	6.3 _{1.1} / 22.1 _{1.1} / 32.9 _{1.1}	6.3 _{1.1} / 15.9 _{1.1} / 41.6 _{1.1}	6.3 _{1.1} / 5.4 _{1.1} / 9.8 _{1.1}	4.4 _{1.1} / 8.9 _{1.1} / 16.9 _{1.1}
CoMML LLM	GCN-DE	5.4 _{1.1} / 5.3 _{1.1} / 61.0 _{1.1}	5.6 _{1.1} / 6.3 _{1.1} / 58.3 _{1.1}	n.a. / n.a. / n.a.	8.0 _{1.1} / 21.5 _{1.1} / 25.1 _{1.1}	8.0 _{1.1} / 16.1 _{1.1} / 42.8 _{1.1}	7.8 _{1.1} / 6.7 _{1.1} / 9.9 _{1.1}	8.3 _{1.1} / 12.2 _{1.1} / 8.5 _{1.1}
	GCN-Ens	4.1 _{1.1} / 4.3 _{1.1} / 62.1 _{1.1}	4.0 _{1.1} / 5.5 _{1.1} / 60.0 _{1.1}	n.a. / n.a. / n.a.	6.4 _{1.1} / 22.8 _{1.1} / 34.0 _{1.1}	6.4 _{1.1} / 13.2 _{1.1} / 38.3 _{1.1}	6.7 _{1.1} / 5.4 _{1.1} / 9.8 _{1.1}	4.6 _{1.1} / 8.9 _{1.1} / 7.4 _{1.1}
	GCN-MCD+DE	5.3 _{1.1} / 5.5 _{1.1} / 61.0 _{1.1}	5.6 _{1.1} / 6.1 _{1.1} / 58.9 _{1.1}	n.a. / n.a. / n.a.	8.0 _{1.1} / 21.5 _{1.1} / 26.3 _{1.1}	8.0 _{1.1} / 16.3 _{1.1} / 41.5 _{1.1}	7.6 _{1.1} / 7.2 _{1.1} / 10.6 _{1.1}	7.8 _{1.1} / 11.1 _{1.1} / 7.9 _{1.1}
	SGCN	23.3 _{1.1} / 24.5 _{1.1} / 40.6 _{1.1}	27.3 _{1.1} / 28.5 _{1.1} / 39.5 _{1.1}	n.a. / n.a. / n.a.	35.3 _{1.1} / 47.0 _{1.1} / 46.5 _{1.1}	35.3 _{1.1} / 19.6 _{1.1} / 12.2 _{1.1}	38.3 _{1.1} / 35.4 _{1.1} / 43.3 _{1.1}	40.5 _{1.1} / 48.0 _{1.1} / 33.1 _{1.1}
	SPN	5.5 _{1.1} / 6.3 _{1.1} / 59.4 _{1.1}	7.8 _{1.1} / 8.8 _{1.1} / 54.2 _{1.1}	n.a. / n.a. / n.a.	7.0 _{1.1} / 12.7 _{1.1} / 15.4 _{1.1}	7.0 _{1.1} / 10.4 _{1.1} / 11.1 _{1.1}	9.5 _{1.1} / 8.9 _{1.1} / 8.9 _{1.1}	4.3 _{1.1} / 9.6 _{1.1} / 10.4 _{1.1}
	GCN (Ours)	3.9 _{1.1} / 4.2 _{1.1} / 65.0 _{1.1}	4.1 _{1.1} / 5.3 _{1.1} / 60.0 _{1.1}	n.a. / n.a. / n.a.	6.2 _{1.1} / 23.1 _{1.1} / 33.0 _{1.1}	6.2 _{1.1} / 16.5 _{1.1} / 41.0 _{1.1}	6.1 _{1.1} / 5.6 _{1.1} / 9.9 _{1.1}	4.3 _{1.1} / 8.9 _{1.1} / 7.5 _{1.1}
	BGCN	7.3 _{1.1} / 7.3 _{1.1} / 65.3 _{1.1}	4.1 _{1.1} / 4.1 _{1.1} / 66.0 _{1.1}	10.8 _{1.1} / 16.0 _{1.1} / 12.8 _{1.1}	10.8 _{1.1} / 11.7 _{1.1} / 30.2 _{1.1}	10.8 _{1.1} / 35.3 _{1.1} / 58.5 _{1.1}	10.0 _{1.1} / 8.6 _{1.1} / 14.9 _{1.1}	8.0 _{1.1} / 4.6 _{1.1} / 5.8 _{1.1}
	GCN-MCD	4.3 _{1.1} / 4.2 _{1.1} / 54.3 _{1.1}	4.2 _{1.1} / 4.4 _{1.1} / 55.1 _{1.1}	7.1 _{1.1} / 18.7 _{1.1} / 13.9 _{1.1}	7.1 _{1.1} / 12.3 _{1.1} / 13.2 _{1.1}	7.1 _{1.1} / 34.3 _{1.1} / 63.9 _{1.1}	5.0 _{1.1} / 4.9 _{1.1} / 13.1 _{1.1}	6.4 _{1.1} / 12.8 _{1.1} / 10.4 _{1.1}
	GCN-DE	7.3 _{1.1} / 7.4 _{1.1} / 51.5 _{1.1}	8.7 _{1.1} / 8.8 _{1.1} / 51.9 _{1.1}	10.9 _{1.1} / 16.7 _{1.1} / 7.6 _{1.1}	10.9 _{1.1} / 13.9 _{1.1} / 12.0 _{1.1}	10.9 _{1.1} / 25.4 _{1.1} / 57.4 _{1.1}	8.4 _{1.1} / 9.1 _{1.1} / 15.2 _{1.1}	10.3 _{1.1} / 16.2 _{1.1} / 13.8 _{1.1}
	GCN-Ens	4.1 _{1.1} / 4.1 _{1.1} / 54.2 _{1.1}	4.4 _{1.1} / 4.6 _{1.1} / 54.9 _{1.1}	7.1 _{1.1} / 19.8 _{1.1} / 15.0 _{1.1}	7.1 _{1.1} / 15.8 _{1.1} / 10.6 _{1.1}	7.1 _{1.1} / 32.3 _{1.1} / 59.5 _{1.1}	5.2 _{1.1} / 4.7 _{1.1} / 13.0 _{1.1}	6.8 _{1.1} / 13.0 _{1.1} / 10.8 _{1.1}
Chiswer	GCN-MCD+DE	7.2 _{1.1} / 7.6 _{1.1} / 51.5 _{1.1}	9.1 _{1.1} / 8.9 _{1.1} / 51.7 _{1.1}	10.9 _{1.1} / 16.8 _{1.1} / 8.6 _{1.1}	10.9 _{1.1} / 15.0 _{1.1} / 12.0 _{1.1}	10.9 _{1.1} / 25.1 _{1.1} / 57.1 _{1.1}	8.1 _{1.1} / 8.7 _{1.1} / 15.6 _{1.1}	10.4 _{1.1} / 15.5 _{1.1} / 13.6 _{1.1}
	SGCN	27.1 _{1.1} / 27.4 _{1.1} / 35.9 _{1.1}	31.5 _{1.1} / 31.9 _{1.1} / 35.4 _{1.1}	37.0 _{1.1} / 45.2 _{1.1} / 32.2 _{1.1}	37.0 _{1.1} / 43.9 _{1.1} / 28.6 _{1.1}	37.0 _{1.1} / 21.2 _{1.1} / 46.3 _{1.1}	35.7 _{1.1} / 35.9 _{1.1} / 47.1 _{1.1}	36.0 _{1.1} / 47.3 _{1.1} / 45.3 _{1.1}
	SPN	4.4 _{1.1} / 4.4 _{1.1} / 57.6 _{1.1}	3.4 _{1.1} / 3.4 _{1.1} / 62.1 _{1.1}	6.1 _{1.1} / 11.6 _{1.1} / 9.8 _{1.1}	6.1 _{1.1} / 11.0 _{1.1} / 12.8 _{1.1}	6.1 _{1.1} / 7.2 _{1.1} / 6.8 _{1.1}	6.3 _{1.1} / 6.1 _{1.1} / 6.8 _{1.1}	5.6 _{1.1} / 6.3 _{1.1} / 6.2 _{1.1}
	GCN (Ours)	4.1 _{1.1} / 4.0 _{1.1} / 54.4 _{1.1}	4.4 _{1.1} / 4.6 _{1.1} / 54.9 _{1.1}	7.3 _{1.1} / 19.3 _{1.1} / 13.7 _{1.1}	7.3 _{1.1} / 13.9 _{1.1} / 13.4 _{1.1}	7.3 _{1.1} / 35.9 _{1.1} / 64.0 _{1.1}	5.4 _{1.1} / 5.3 _{1.1} / 12.8 _{1.1}	6.1 _{1.1} / 12.8 _{1.1} / 10.7 _{1.1}
PubMed	BGCN	4.8 _{1.1} / 5.0 _{1.1} / 93.0 _{1.1}	8.9 _{1.1} / 8.6 _{1.1} / 81.9 _{1.1}	n.a. / n.a. / n.a.	4.3 _{1.1} / 8.3 _{1.1} / 12.0 _{1.1}	4.3 _{1.1} / 27.1 _{1.1} / 47.9 _{1.1}	11.9 _{1.1} / 12.8 _{1.1} / 9.5 _{1.1}	9.4 _{1.1} / 4.3 _{1.1} / 19.7 _{1.1}
	GCN-MCD	2.0 _{1.1} / 2.6 _{1.1} / 83.8 _{1.1}	2.7 _{1.1} / 2.8 _{1.1} / 71.4 _{1.1}	n.a. / n.a. / n.a.	4.0 _{1.1} / 12.9 _{1.1} / 7.7 _{1.1}	4.0 _{1.1} / 23.4 _{1.1} / 46.9 _{1.1}	2.6 _{1.1} / 3.1 _{1.1} / 6.7 _{1.1}	2.9 _{1.1} / 10.9 _{1.1} / 10.4 _{1.1}
	GCN-DE	2.7 _{1.1} / 3.3 _{1.1} / 82.7 _{1.1}	2.6 _{1.1} / 2.8 _{1.1} / 70.5 _{1.1}	n.a. / n.a. / n.a.	5.4 _{1.1} / 9.4 _{1.1} / 8.0 _{1.1}	5.4 _{1.1} / 21.7 _{1.1} / 48.7 _{1.1}	4.1 _{1.1} / 3.5 _{1.1} / 6.0 _{1.1}	5.3 _{1.1} / 11.0 _{1.1} / 4.6 _{1.1}
	GCN-Ens	1.8 _{1.1} / 2.5 _{1.1} / 83.5 _{1.1}	2.7 _{1.1} / 2.8 _{1.1} / 71.3 _{1.1}	n.a. / n.a. / n.a.	3.9 _{1.1} / 13.1 _{1.1} / 8.2 _{1.1}	3.9 _{1.1} / 22.7 _{1.1} / 45.5 _{1.1}	2.7 _{1.1} / 3.1 _{1.1} / 6.7 _{1.1}	2.8 _{1.1} / 11.1 _{1.1} / 5.8 _{1.1}
	GCN-MCD+DE	2.6 _{1.1} / 3.0 _{1.1} / 82.9 _{1.1}	8.5 _{1.1} / 2.8 _{1.1} / 70.7 _{1.1}	n.a. / n.a. / n.a.	5.9 _{1.1} / 10.4 _{1.1} / 10.3 _{1.1}	5.9 _{1.1} / 20.7 _{1.1} / 47.8 _{1.1}	4.4 _{1.1} / 4.0 _{1.1} / 6.5 _{1.1}	5.1 _{1.1} / 10.5 _{1.1} / 4.7 _{1.1}
	SGCN	9.8 _{1.1} / 11.0 _{1.1} / 75.3 _{1.1}	8.5 _{1.1} / 9.1 _{1.1} / 67.1 _{1.1}	n.a. / n.a. / n.a.	16.5 _{1.1} / 25.7 _{1.1} / 14.9 _{1.1}	16.5 _{1.1} / 10.8 _{1.1} / 35.3 _{1.1}	16.4 _{1.1} / 13.6 _{1.1} / 23.1 _{1.1}	19.1 _{1.1} / 30.6 _{1.1} / 12.0 _{1.1}
	SPN	3.5 _{1.1} / 5.1 _{1.1} / 82.1 _{1.1}	2.0 _{1.1} / 2.4 _{1.1} / 74.3 _{1.1}	n.a. / n.a. / n.a.	3.8 _{1.1} / 12.0 _{1.1} / 9.1 _{1.1}	3.8 _{1.1} / 8.3 _{1.1} / 8.4 _{1.1}	4.1 _{1.1} / 3.7 _{1.1} / 5.3 _{1.1}	3.8 _{1.1} / 5.9 _{1.1} / 14.4 _{1.1}
	GCN (Ours)	2.9 _{1.1} / 2.5 _{1.1} / 83.4 _{1.1}	2.7 _{1.1} / 2.8 _{1.1} / 71.4 _{1.1}	n.a. / n.a. / n.a.	3.9 _{1.1} / 12.8 _{1.1} / 7.6 _{1.1}	3.9 _{1.1} / 22.4 _{1.1} / 46.9 _{1.1}	2.8 _{1.1} / 3.3 _{1.1} / 6.8 _{1.1}	2.7 _{1.1} / 11.1 _{1.1} / 16.1 _{1.1}
	BGCN	4.0 _{1.1} / 3.9 _{1.1} / 86.7 _{1.1}	2.5 _{1.1} / 2.4 _{1.1} / 82.6 _{1.1}	4.4 _{1.1} / 8.7 _{1.1} / 9.4 _{1.1}	4.4 _{1.1} / 23.7 _{1.1} / 30.0 _{1.1}	4.4 _{1.1} / 31.5 _{1.1} / 39.4 _{1.1}	4.3 _{1.1} / 3.8 _{1.1} / 4.5 _{1.1}	1.6 _{1.1} / 2.8 _{1.1} / 6.6 _{1.1}
	GCN-MCD	3.8 _{1.1} / 4.9 _{1.1} / 73.1 _{1.1}	4.4 _{1.1} / 5.8 _{1.1} / 70.9 _{1.1}	6.4 _{1.1} / 25.9 _{1.1} / 26.4 _{1.1}	6.4 _{1.1} / 11.3 _{1.1} / 12.4 _{1.1}	6.4 _{1.1} / 22.9 _{1.1} / 35.1 _{1.1}	4.5 _{1.1} / 3.0 _{1.1} / 13.4 _{1.1}	4.6 _{1.1} / 13.1 _{1.1} / 10.1 _{1.1}
	GCN-DE	5.4 _{1.1} / 6.7 _{1.1} / 69.4 _{1.1}	5.8 _{1.1} / 7.1 _{1.1} / 67.7 _{1.1}	9.0 _{1.1} / 29.1 _{1.1} / 28.0 _{1.1}	9.0 _{1.1} / 13.7 _{1.1} / 11.0 _{1.1}	9.0 _{1.1} / 20.7 _{1.1} / 32.6 _{1.1}	6.4 _{1.1} / 4.3 _{1.1} / 14.9 _{1.1}	9.1 _{1.1} / 17.3 _{1.1} / 13.4 _{1.1}
Amazon Photo	GCN-Ens	3.7 _{1.1} / 4.9 _{1.1} / 73.0 _{1.1}	4.2 _{1.1} / 5.6 _{1.1} / 71.2 _{1.1}	6.4 _{1.1} / 26.0 _{1.1} / 27.1 _{1.1}	6.4 _{1.1} / 11.9 _{1.1} / 12.5 _{1.1}	6.4 _{1.1} / 19.2 _{1.1} / 30.9 _{1.1}	4.6 _{1.1} / 3.0 _{1.1} / 13.5 _{1.1}	4.7 _{1.1} / 13.1 _{1.1} / 10.1 _{1.1}
	GCN-MCD+DE	5.4 _{1.1} / 6.5 _{1.1} / 69.4 _{1.1}	5.9 _{1.1} / 7.3 _{1.1} / 67.8 _{1.1}	9.8 _{1.1} / 29.8 _{1.1} / 28.6 _{1.1}	9.8 _{1.1} / 13.6 _{1.1} / 11.1 _{1.1}	9.8 _{1.1} / 20.5 _{1.1} / 32.6 _{1.1}	7.1 _{1.1} / 4.8 _{1.1} / 15.5 _{1.1}	8.8 _{1.1}

	Model	LoC (i.i.d.) Clean / I.d. / O.o.D.	LoC (heterophilic) Clean / I.d. / O.o.D.	Ber(β) (near) Clean / I.d. / O.o.D.	Ber(0.5) Clean / I.d. / O.o.D.	$N(0, 1)$ (far) Clean / I.d. / O.o.D.	Page Rank Clean / I.d. / O.o.D.	Homophily Clean / I.d. / O.o.D.
CoMML	BGCN	6.0 _{1.1} / 6.7 _{1.1} / 75.5 _{1.1}	9.5 _{1.1} / 9.2 _{1.1} / 75.1 _{1.1}	9.7 _{1.1} / 8.3 _{1.1} / 8.9 _{1.1}	9.7 _{1.1} / 11.7 _{1.1} / 20.1 _{1.1}	9.7 _{1.1} / 53.0 _{1.1} / 61.9 _{1.1}	11.5 _{1.1} / 11.4 _{1.1} / 8.9 _{1.1}	3.6 _{1.1} / 3.6 _{1.1} / 16.6 _{1.1}
	GCN-MCD	6.0 _{1.1} / 6.0 _{1.1} / 56.9 _{1.1}	7.0 _{1.1} / 7.0 _{1.1} / 54.9 _{1.1}	8.4 _{1.1} / 20.3 _{1.1} / 22.3 _{1.1}	8.4 _{1.1} / 20.0 _{1.1} / 16.8 _{1.1}	8.4 _{1.1} / 38.7 _{1.1} / 67.1 _{1.1}	7.3 _{1.1} / 7.3 _{1.1} / 14.8 _{1.1}	9.3 _{1.1} / 9.3 _{1.1} / 7.0 _{1.1}
	GCN-DE	11.7 _{1.1} / 11.5 _{1.1} / 50.7 _{1.1}	10.6 _{1.1} / 10.6 _{1.1} / 49.2 _{1.1}	15.1 _{1.1} / 24.3 _{1.1} / 17.9 _{1.1}	15.1 _{1.1} / 21.8 _{1.1} / 14.6 _{1.1}	15.1 _{1.1} / 38.3 _{1.1} / 59.5 _{1.1}	12.1 _{1.1} / 12.6 _{1.1} / 19.3 _{1.1}	19.2 _{1.1} / 19.7 _{1.1} / 11.6 _{1.1}
	GCN-Ens	5.7 _{1.1} / 5.7 _{1.1} / 56.8 _{1.1}	7.1 _{1.1} / 7.1 _{1.1} / 54.9 _{1.1}	9.0 _{1.1} / 21.1 _{1.1} / 23.0 _{1.1}	9.0 _{1.1} / 21.2 _{1.1} / 17.9 _{1.1}	9.0 _{1.1} / 45.5 _{1.1} / 63.8 _{1.1}	7.4 _{1.1} / 7.4 _{1.1} / 14.7 _{1.1}	9.4 _{1.1} / 9.4 _{1.1} / 7.1 _{1.1}
	GCN-MCD+DE	11.0 _{1.1} / 11.0 _{1.1} / 51.0 _{1.1}	9.7 _{1.1} / 10.3 _{1.1} / 50.2 _{1.1}	15.0 _{1.1} / 23.7 _{1.1} / 18.3 _{1.1}	15.0 _{1.1} / 21.0 _{1.1} / 13.5 _{1.1}	15.0 _{1.1} / 36.7 _{1.1} / 57.3 _{1.1}	11.4 _{1.1} / 11.9 _{1.1} / 17.0 _{1.1}	19.4 _{1.1} / 18.7 _{1.1} / 12.0 _{1.1}
	SGCN	28.5 _{1.1} / 28.5 _{1.1} / 38.5 _{1.1}	26.9 _{1.1} / 26.9 _{1.1} / 39.5 _{1.1}	27.4 _{1.1} / 29.9 _{1.1} / 24.6 _{1.1}	27.4 _{1.1} / 27.7 _{1.1} / 21.0 _{1.1}	27.4 _{1.1} / 11.3 _{1.1} / 23.0 _{1.1}	40.5 _{1.1} / 40.5 _{1.1} / 40.7 _{1.1}	49.6 _{1.1} / 49.6 _{1.1} / 38.0 _{1.1}
LLaMA	GCN	7.6 _{1.1} / 7.6 _{1.1} / 56.5 _{1.1}	10.3 _{1.1} / 10.3 _{1.1} / 54.4 _{1.1}	12.2 _{1.1} / 14.1 _{1.1} / 13.0 _{1.1}	12.2 _{1.1} / 15.1 _{1.1} / 13.2 _{1.1}	12.2 _{1.1} / 12.4 _{1.1} / 10.6 _{1.1}	11.3 _{1.1} / 11.3 _{1.1} / 10.9 _{1.1}	10.0 _{1.1} / 10.0 _{1.1} / 8.6 _{1.1}
	GCN (Ours)	5.9 _{1.1} / 5.9 _{1.1} / 56.8 _{1.1}	6.9 _{1.1} / 6.9 _{1.1} / 54.9 _{1.1}	8.5 _{1.1} / 20.4 _{1.1} / 22.3 _{1.1}	8.5 _{1.1} / 20.5 _{1.1} / 18.3 _{1.1}	8.5 _{1.1} / 49.8 _{1.1} / 68.6 _{1.1}	7.6 _{1.1} / 7.6 _{1.1} / 14.5 _{1.1}	9.6 _{1.1} / 9.6 _{1.1} / 7.5 _{1.1}
	BGCN	3.9 _{1.1} / 4.0 _{1.1} / 71.0 _{1.1}	4.8 _{1.1} / 4.7 _{1.1} / 70.8 _{1.1}	n.a. / n.a. / n.a.	6.4 _{1.1} / 14.8 _{1.1} / 24.0 _{1.1}	6.4 _{1.1} / 21.6 _{1.1} / 33.7 _{1.1}	6.0 _{1.1} / 6.1 _{1.1} / 7.6 _{1.1}	3.2 _{1.1} / 3.3 _{1.1} / 15.3 _{1.1}
	GCN-MCD	4.5 _{1.1} / 4.5 _{1.1} / 64.4 _{1.1}	4.3 _{1.1} / 4.3 _{1.1} / 65.3 _{1.1}	n.a. / n.a. / n.a.	6.4 _{1.1} / 17.5 _{1.1} / 25.3 _{1.1}	6.4 _{1.1} / 20.2 _{1.1} / 45.7 _{1.1}	5.8 _{1.1} / 5.8 _{1.1} / 10.6 _{1.1}	5.0 _{1.1} / 5.0 _{1.1} / 10.8 _{1.1}
	GCN-DE	4.9 _{1.1} / 5.2 _{1.1} / 61.4 _{1.1}	5.5 _{1.1} / 5.3 _{1.1} / 62.0 _{1.1}	n.a. / n.a. / n.a.	8.1 _{1.1} / 18.4 _{1.1} / 22.7 _{1.1}	8.1 _{1.1} / 17.9 _{1.1} / 41.7 _{1.1}	7.9 _{1.1} / 8.0 _{1.1} / 11.6 _{1.1}	7.7 _{1.1} / 8.0 _{1.1} / 8.6 _{1.1}
	GCN-Ens	4.2 _{1.1} / 4.2 _{1.1} / 64.4 _{1.1}	4.4 _{1.1} / 4.4 _{1.1} / 65.3 _{1.1}	n.a. / n.a. / n.a.	6.8 _{1.1} / 18.0 _{1.1} / 26.3 _{1.1}	6.8 _{1.1} / 17.4 _{1.1} / 45.8 _{1.1}	5.6 _{1.1} / 5.6 _{1.1} / 10.9 _{1.1}	4.9 _{1.1} / 4.9 _{1.1} / 10.5 _{1.1}
Amazon	GCN-MCD+DE	5.1 _{1.1} / 5.0 _{1.1} / 61.4 _{1.1}	5.6 _{1.1} / 5.6 _{1.1} / 61.7 _{1.1}	n.a. / n.a. / n.a.	8.4 _{1.1} / 18.0 _{1.1} / 22.3 _{1.1}	8.4 _{1.1} / 17.8 _{1.1} / 42.9 _{1.1}	7.7 _{1.1} / 7.7 _{1.1} / 10.9 _{1.1}	7.8 _{1.1} / 8.0 _{1.1} / 8.7 _{1.1}
	SGCN	21.1 _{1.1} / 21.1 _{1.1} / 41.1 _{1.1}	24.5 _{1.1} / 24.5 _{1.1} / 41.7 _{1.1}	n.a. / n.a. / n.a.	36.3 _{1.1} / 43.3 _{1.1} / 41.6 _{1.1}	36.3 _{1.1} / 17.8 _{1.1} / 12.2 _{1.1}	35.4 _{1.1} / 35.4 _{1.1} / 45.9 _{1.1}	42.6 _{1.1} / 42.6 _{1.1} / 30.6 _{1.1}
	GCN	6.4 _{1.1} / 6.4 _{1.1} / 56.5 _{1.1}	8.1 _{1.1} / 8.1 _{1.1} / 55.7 _{1.1}	n.a. / n.a. / n.a.	8.9 _{1.1} / 10.0 _{1.1} / 11.3 _{1.1}	8.9 _{1.1} / 9.5 _{1.1} / 19.3 _{1.1}	9.2 _{1.1} / 9.2 _{1.1} / 8.8 _{1.1}	9.7 _{1.1} / 9.7 _{1.1} / 8.8 _{1.1}
	GCN (Ours)	4.2 _{1.1} / 4.2 _{1.1} / 64.4 _{1.1}	4.6 _{1.1} / 4.6 _{1.1} / 65.0 _{1.1}	n.a. / n.a. / n.a.	6.7 _{1.1} / 18.2 _{1.1} / 27.0 _{1.1}	6.7 _{1.1} / 18.8 _{1.1} / 43.2 _{1.1}	5.9 _{1.1} / 5.9 _{1.1} / 10.5 _{1.1}	5.0 _{1.1} / 5.0 _{1.1} / 10.7 _{1.1}
	BGCN	8.4 _{1.1} / 8.4 _{1.1} / 66.4 _{1.1}	4.6 _{1.1} / 3.9 _{1.1} / 71.1 _{1.1}	8.6 _{1.1} / 8.5 _{1.1} / 12.5 _{1.1}	8.6 _{1.1} / 13.3 _{1.1} / 20.0 _{1.1}	8.6 _{1.1} / 37.8 _{1.1} / 53.2 _{1.1}	8.0 _{1.1} / 7.3 _{1.1} / 5.4 _{1.1}	5.4 _{1.1} / 5.4 _{1.1} / 8.5 _{1.1}
	GCN-MCD	3.8 _{1.1} / 3.8 _{1.1} / 54.1 _{1.1}	5.2 _{1.1} / 5.2 _{1.1} / 58.6 _{1.1}	5.6 _{1.1} / 9.2 _{1.1} / 8.1 _{1.1}	5.6 _{1.1} / 8.4 _{1.1} / 10.4 _{1.1}	5.6 _{1.1} / 40.5 _{1.1} / 68.9 _{1.1}	5.3 _{1.1} / 5.3 _{1.1} / 12.3 _{1.1}	5.1 _{1.1} / 5.1 _{1.1} / 5.7 _{1.1}
PubMed	GCN-DE	6.0 _{1.1} / 7.1 _{1.1} / 50.2 _{1.1}	9.2 _{1.1} / 8.9 _{1.1} / 53.7 _{1.1}	9.3 _{1.1} / 10.8 _{1.1} / 8.0 _{1.1}	9.3 _{1.1} / 9.8 _{1.1} / 14.0 _{1.1}	9.3 _{1.1} / 28.4 _{1.1} / 58.3 _{1.1}	10.3 _{1.1} / 10.1 _{1.1} / 16.2 _{1.1}	10.0 _{1.1} / 10.1 _{1.1} / 7.7 _{1.1}
	GCN-Ens	3.9 _{1.1} / 3.9 _{1.1} / 53.0 _{1.1}	5.2 _{1.1} / 5.2 _{1.1} / 58.4 _{1.1}	6.0 _{1.1} / 9.0 _{1.1} / 8.8 _{1.1}	6.0 _{1.1} / 9.4 _{1.1} / 11.5 _{1.1}	6.0 _{1.1} / 38.3 _{1.1} / 63.3 _{1.1}	5.0 _{1.1} / 5.0 _{1.1} / 12.0 _{1.1}	5.0 _{1.1} / 5.0 _{1.1} / 5.7 _{1.1}
	GCN-MCD+DE	6.8 _{1.1} / 6.8 _{1.1} / 50.9 _{1.1}	8.6 _{1.1} / 8.7 _{1.1} / 54.2 _{1.1}	8.9 _{1.1} / 10.3 _{1.1} / 8.1 _{1.1}	8.9 _{1.1} / 9.3 _{1.1} / 12.8 _{1.1}	8.9 _{1.1} / 20.2 _{1.1} / 59.6 _{1.1}	9.9 _{1.1} / 10.0 _{1.1} / 15.7 _{1.1}	10.3 _{1.1} / 10.2 _{1.1} / 8.2 _{1.1}
	SGCN	25.0 _{1.1} / 25.0 _{1.1} / 36.1 _{1.1}	29.6 _{1.1} / 29.6 _{1.1} / 37.0 _{1.1}	38.0 _{1.1} / 40.8 _{1.1} / 30.2 _{1.1}	38.0 _{1.1} / 40.8 _{1.1} / 29.1 _{1.1}	38.0 _{1.1} / 23.2 _{1.1} / 52.0 _{1.1}	35.8 _{1.1} / 35.8 _{1.1} / 47.9 _{1.1}	39.2 _{1.1} / 39.2 _{1.1} / 37.2 _{1.1}
	GCN	4.8 _{1.1} / 4.8 _{1.1} / 61.4 _{1.1}	3.2 _{1.1} / 3.2 _{1.1} / 66.7 _{1.1}	4.8 _{1.1} / 7.0 _{1.1} / 7.8 _{1.1}	4.8 _{1.1} / 6.4 _{1.1} / 10.5 _{1.1}	4.8 _{1.1} / 5.4 _{1.1} / 6.0 _{1.1}	5.1 _{1.1} / 5.1 _{1.1} / 4.8 _{1.1}	4.7 _{1.1} / 4.7 _{1.1} / 5.1 _{1.1}
	GCN (Ours)	4.1 _{1.1} / 4.1 _{1.1} / 54.2 _{1.1}	5.2 _{1.1} / 5.2 _{1.1} / 58.8 _{1.1}	5.5 _{1.1} / 9.2 _{1.1} / 8.4 _{1.1}	5.5 _{1.1} / 6.2 _{1.1} / 10.8 _{1.1}	5.5 _{1.1} / 41.8 _{1.1} / 67.7 _{1.1}	5.3 _{1.1} / 5.3 _{1.1} / 12.1 _{1.1}	5.3 _{1.1} / 5.3 _{1.1} / 5.4 _{1.1}
Amazon Photo	BGCN	5.7 _{1.1} / 5.7 _{1.1} / 92.5 _{1.1}	10.2 _{1.1} / 10.2 _{1.1} / 90.2 _{1.1}	n.a. / n.a. / n.a.	10.2 _{1.1} / 10.0 _{1.1} / 16.7 _{1.1}	10.2 _{1.1} / 37.0 _{1.1} / 52.5 _{1.1}	12.4 _{1.1} / 12.4 _{1.1} / 8.4 _{1.1}	5.0 _{1.1} / 5.7 _{1.1} / 21.3 _{1.1}
	GCN-MCD	2.4 _{1.1} / 2.4 _{1.1} / 83.3 _{1.1}	2.2 _{1.1} / 2.2 _{1.1} / 77.2 _{1.1}	n.a. / n.a. / n.a.	3.8 _{1.1} / 9.2 _{1.1} / 8.7 _{1.1}	3.8 _{1.1} / 28.5 _{1.1} / 49.0 _{1.1}	3.3 _{1.1} / 3.3 _{1.1} / 8.6 _{1.1}	2.7 _{1.1} / 2.7 _{1.1} / 14.2 _{1.1}
	GCN-DE	3.2 _{1.1} / 3.1 _{1.1} / 82.0 _{1.1}	2.3 _{1.1} / 2.4 _{1.1} / 73.3 _{1.1}	n.a. / n.a. / n.a.	4.5 _{1.1} / 7.7 _{1.1} / 6.1 _{1.1}	4.5 _{1.1} / 24.8 _{1.1} / 50.2 _{1.1}	4.3 _{1.1} / 4.2 _{1.1} / 7.9 _{1.1}	4.6 _{1.1} / 4.5 _{1.1} / 9.6 _{1.1}
	GCN-Ens	2.3 _{1.1} / 2.3 _{1.1} / 83.3 _{1.1}	2.2 _{1.1} / 2.2 _{1.1} / 77.4 _{1.1}	n.a. / n.a. / n.a.	3.7 _{1.1} / 8.9 _{1.1} / 8.4 _{1.1}	3.7 _{1.1} / 29.6 _{1.1} / 49.4 _{1.1}	3.1 _{1.1} / 3.1 _{1.1} / 14.2 _{1.1}	2.6 _{1.1} / 2.6 _{1.1} / 14.2 _{1.1}
	GCN-MCD+DE	3.4 _{1.1} / 3.5 _{1.1} / 80.8 _{1.1}	2.3 _{1.1} / 2.1 _{1.1} / 73.3 _{1.1}	n.a. / n.a. / n.a.	5.0 _{1.1} / 8.3 _{1.1} / 5.3 _{1.1}	5.0 _{1.1} / 24.0 _{1.1} / 49.4 _{1.1}	3.9 _{1.1} / 4.1 _{1.1} / 7.5 _{1.1}	4.3 _{1.1} / 4.3 _{1.1} / 9.9 _{1.1}
	SGCN	10.3 _{1.1} / 10.3 _{1.1} / 74.9 _{1.1}	8.4 _{1.1} / 8.4 _{1.1} / 69.6 _{1.1}	n.a. / n.a. / n.a.	18.1 _{1.1} / 24.4 _{1.1} / 18.6 _{1.1}	18.1 _{1.1} / 14.8 _{1.1} / 39.6 _{1.1}	14.6 _{1.1} / 14.6 _{1.1} / 24.4 _{1.1}	22.1 _{1.1} / 22.1 _{1.1} / 6.3 _{1.1}
Amazon Computers	GCN	4.5 _{1.1} / 4.5 _{1.1} / 80.5 _{1.1}	3.0 _{1.1} / 3.0 _{1.1} / 78.2 _{1.1}	n.a. / n.a. / n.a.	4.1 _{1.1} / 6.9 _{1.1} / 8.7 _{1.1}	4.1 _{1.1} / 4.8 _{1.1} / 4.9 _{1.1}	4.5 _{1.1} / 4.5 _{1.1} / 3.8 _{1.1}	4.8 _{1.1} / 4.8 _{1.1} / 11.2 _{1.1}
	GCN (Ours)	2.3 _{1.1} / 2.3 _{1.1} / 83.3 _{1.1}	2.3 _{1.1} / 2.3 _{1.1} / 77.5 _{1.1}	n.a. / n.a. / n.a.	3.9 _{1.1} / 8.8 _{1.1} / 18.4 _{1.1}	3.9 _{1.1} / 20.4 _{1.1} / 50.2 _{1.1}	3.3 _{1.1} / 3.3 _{1.1} / 9.4 _{1.1}	2.8 _{1.1} / 2.8 _{1.1} / 14.3 _{1.1}
	BGCN	3.5 _{1.1} / 3.5 _{1.1} / 84.1 _{1.1}	1.9 _{1.1} / 1.9 _{1.1} / 84.9 _{1.1}	4.2 _{1.1} / 5.7 _{1.1} / 6.1 _{1.1}	4.2 _{1.1} / 23.6 _{1.1} / 31.9 _{1.1}	4.2 _{1.1} / 27.6 _{1.1} / 37.4 _{1.1}	3.3 _{1.1} / 3.5 _{1.1} / 5.8 _{1.1}	1.3 _{1.1} / 1.3 _{1.1} / 10.8 _{1.1}
	GCN-MCD	3.6 _{1.1} / 3.6 _{1.1} / 74.8 _{1.1}	4.2 _{1.1} / 4.2 _{1.1} / 72.2 _{1.1}	5.7 _{1.1} / 21.7 _{1.1} / 22.7 _{1.1}	5.7 _{1.1} / 9.8 _{1.1} / 15.8 _{1.1}	5.7 _{1.1} / 21.1 _{1.1} / 33.0 _{1.1}	4.0 _{1.1} / 4.0 _{1.1} / 17.6 _{1.1}	5.0 _{1.1} / 5.0 _{1.1} / 5.0 _{1.1}
	GCN-DE	5.1 _{1.1} / 4.8 _{1.1} / 71.5 _{1.1}	6.4 _{1.1} / 6.2 _{1.1} / 69.1 _{1.1}	8.6 _{1.1} / 25.5 _{1.1} / 25.4 _{1.1}	8.6 _{1.1} / 8.9 _{1.1} / 11.4 _{1.1}	8.6 _{1.1} / 21.8 _{1.1} / 33.7 _{1.1}	6.4 _{1.1} / 6.4 _{1.1} / 18.9 _{1.1}	9.0 _{1.1} / 8.9 _{1.1} / 5.1 _{1.1}
	GCN-Ens	3.6 _{1.1} / 3.6 _{1.1} / 74.9 _{1.1}	4.4 _{1.1} / 4.4 _{1.1} / 71.8 _{1.1}	5.9 _{1.1} / 22.0 _{1.1} / 23.1 _{1.1}	5.9 _{1.1} / 9.3 _{1.1} / 16.0 _{1.1}	5.9 _{1.1} / 20.2 _{1.1} / 33.1 _{1.1}	4.3 _{1.1} / 4.3 _{1.1} / 17.7 _{1.1}	5.6 _{1.1} / 5.6 _{1.1} / 4.9 _{1.1}

	Model	LoC (last) Clean / I.d. / O.o.D.	LoC (heterophily) Clean / I.d. / O.o.D.	Ber(p) (near) Clean / I.d. / O.o.D.	Ber(0.5) Clean / I.d. / O.o.D.	N(0, 1) (far) Clean / I.d. / O.o.D.	Page Rank Clean / I.d. / O.o.D.	Homophily Clean / I.d. / O.o.D.
CoMML	BGCN	0.17 ₁₀₀ / 0.16 ₁₀₀ / 1.59 ₁₀₀	0.23 ₁₀₀ / 0.23 ₁₀₀ / 1.57 ₁₀₀	0.40 ₁₀₀ / 0.39 ₁₀₀ / 0.47 ₁₀₀	0.40 ₁₀₀ / 0.36 ₁₀₀ / 0.73 ₁₀₀	0.40 ₁₀₀ / 1.18 ₁₀₀ / 1.30 ₁₀₀	0.32 ₁₀₀ / 0.31 ₁₀₀ / 0.32 ₁₀₀	0.17 ₁₀₀ / 0.11 ₁₀₀ / 0.42 ₁₀₀
	GCN-MCD	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.40 ₁₀₀	0.20 ₁₀₀ / 0.21 ₁₀₀ / 1.38 ₁₀₀	0.35 ₁₀₀ / 0.43 ₁₀₀ / 0.53 ₁₀₀	0.35 ₁₀₀ / 0.44 ₁₀₀ / 0.55 ₁₀₀	0.35 ₁₀₀ / 1.09 ₁₀₀ / 1.41 ₁₀₀	0.26 ₁₀₀ / 0.35 ₁₀₀ / 0.31 ₁₀₀	0.17 ₁₀₀ / 0.17 ₁₀₀ / 0.45 ₁₀₀
	GCN-DE	0.21 ₁₀₀ / 0.21 ₁₀₀ / 1.36 ₁₀₀	0.28 ₁₀₀ / 0.29 ₁₀₀ / 1.35 ₁₀₀	0.44 ₁₀₀ / 0.53 ₁₀₀ / 0.66 ₁₀₀	0.44 ₁₀₀ / 0.56 ₁₀₀ / 0.71 ₁₀₀	0.44 ₁₀₀ / 1.02 ₁₀₀ / 1.31 ₁₀₀	0.33 ₁₀₀ / 0.32 ₁₀₀ / 0.40 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 0.55 ₁₀₀
	GCN-Ens	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.39 ₁₀₀	0.20 ₁₀₀ / 0.20 ₁₀₀ / 1.38 ₁₀₀	0.35 ₁₀₀ / 0.42 ₁₀₀ / 0.53 ₁₀₀	0.35 ₁₀₀ / 0.43 ₁₀₀ / 0.55 ₁₀₀	0.35 ₁₀₀ / 1.08 ₁₀₀ / 1.34 ₁₀₀	0.26 ₁₀₀ / 0.25 ₁₀₀ / 0.31 ₁₀₀	0.17 ₁₀₀ / 0.17 ₁₀₀ / 0.45 ₁₀₀
	GCN-MCD+DE	0.21 ₁₀₀ / 0.21 ₁₀₀ / 1.36 ₁₀₀	0.28 ₁₀₀ / 0.29 ₁₀₀ / 1.35 ₁₀₀	0.44 ₁₀₀ / 0.53 ₁₀₀ / 0.66 ₁₀₀	0.44 ₁₀₀ / 0.56 ₁₀₀ / 0.71 ₁₀₀	0.44 ₁₀₀ / 1.07 ₁₀₀ / 1.30 ₁₀₀	0.33 ₁₀₀ / 0.32 ₁₀₀ / 0.40 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 0.54 ₁₀₀
	SGCN	0.31 ₁₀₀ / 0.33 ₁₀₀ / 1.27 ₁₀₀	0.30 ₁₀₀ / 0.32 ₁₀₀ / 1.27 ₁₀₀	0.80 ₁₀₀ / 0.82 ₁₀₀ / 0.84 ₁₀₀	0.80 ₁₀₀ / 0.83 ₁₀₀ / 0.84 ₁₀₀	0.80 ₁₀₀ / 0.86 ₁₀₀ / 0.91 ₁₀₀	0.55 ₁₀₀ / 0.53 ₁₀₀ / 0.65 ₁₀₀	0.53 ₁₀₀ / 0.58 ₁₀₀ / 0.72 ₁₀₀
CoMML LLM	GCN	0.17 ₁₀₀ / 0.17 ₁₀₀ / 1.42 ₁₀₀	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.40 ₁₀₀	0.40 ₁₀₀ / 0.37 ₁₀₀ / 0.38 ₁₀₀	0.40 ₁₀₀ / 0.38 ₁₀₀ / 0.40 ₁₀₀	0.40 ₁₀₀ / 0.36 ₁₀₀ / 0.35 ₁₀₀	0.33 ₁₀₀ / 0.32 ₁₀₀ / 0.33 ₁₀₀	0.17 ₁₀₀ / 0.26 ₁₀₀ / 0.63 ₁₀₀
	GCN (Ours)	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.40 ₁₀₀	0.20 ₁₀₀ / 0.21 ₁₀₀ / 1.39 ₁₀₀	0.35 ₁₀₀ / 0.42 ₁₀₀ / 0.53 ₁₀₀	0.35 ₁₀₀ / 0.43 ₁₀₀ / 0.54 ₁₀₀	0.35 ₁₀₀ / 1.12 ₁₀₀ / 1.40 ₁₀₀	0.26 ₁₀₀ / 0.25 ₁₀₀ / 0.31 ₁₀₀	0.17 ₁₀₀ / 0.17 ₁₀₀ / 0.45 ₁₀₀
	BGCN	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.59 ₁₀₀	0.16 ₁₀₀ / 0.16 ₁₀₀ / 1.54 ₁₀₀	n.a. / n.a. / n.a.	0.32 ₁₀₀ / 0.34 ₁₀₀ / 0.45 ₁₀₀	0.32 ₁₀₀ / 0.72 ₁₀₀ / 1.06 ₁₀₀	0.23 ₁₀₀ / 0.23 ₁₀₀ / 0.26 ₁₀₀	0.14 ₁₀₀ / 0.08 ₁₀₀ / 0.43 ₁₀₀
	GCN-MCD	0.14 ₁₀₀ / 0.14 ₁₀₀ / 1.51 ₁₀₀	0.15 ₁₀₀ / 0.16 ₁₀₀ / 1.49 ₁₀₀	n.a. / n.a. / n.a.	0.32 ₁₀₀ / 0.35 ₁₀₀ / 0.47 ₁₀₀	0.32 ₁₀₀ / 0.74 ₁₀₀ / 1.10 ₁₀₀	0.23 ₁₀₀ / 0.22 ₁₀₀ / 0.27 ₁₀₀	0.14 ₁₀₀ / 0.09 ₁₀₀ / 0.40 ₁₀₀
	GCN-DE	0.17 ₁₀₀ / 0.18 ₁₀₀ / 1.49 ₁₀₀	0.18 ₁₀₀ / 0.18 ₁₀₀ / 1.46 ₁₀₀	n.a. / n.a. / n.a.	0.35 ₁₀₀ / 0.40 ₁₀₀ / 0.55 ₁₀₀	0.35 ₁₀₀ / 0.75 ₁₀₀ / 1.13 ₁₀₀	0.26 ₁₀₀ / 0.25 ₁₀₀ / 0.31 ₁₀₀	0.17 ₁₀₀ / 0.15 ₁₀₀ / 0.43 ₁₀₀
	GCN-Ens	0.14 ₁₀₀ / 0.14 ₁₀₀ / 1.51 ₁₀₀	0.15 ₁₀₀ / 0.16 ₁₀₀ / 1.48 ₁₀₀	n.a. / n.a. / n.a.	0.32 ₁₀₀ / 0.35 ₁₀₀ / 0.47 ₁₀₀	0.32 ₁₀₀ / 0.71 ₁₀₀ / 1.04 ₁₀₀	0.23 ₁₀₀ / 0.22 ₁₀₀ / 0.27 ₁₀₀	0.13 ₁₀₀ / 0.09 ₁₀₀ / 0.40 ₁₀₀
Chester	GCN-MCD+DE	0.17 ₁₀₀ / 0.17 ₁₀₀ / 1.49 ₁₀₀	0.18 ₁₀₀ / 0.19 ₁₀₀ / 1.46 ₁₀₀	n.a. / n.a. / n.a.	0.35 ₁₀₀ / 0.39 ₁₀₀ / 0.55 ₁₀₀	0.35 ₁₀₀ / 0.75 ₁₀₀ / 1.10 ₁₀₀	0.26 ₁₀₀ / 0.25 ₁₀₀ / 0.31 ₁₀₀	0.17 ₁₀₀ / 0.15 ₁₀₀ / 0.43 ₁₀₀
	SGCN	0.24 ₁₀₀ / 0.26 ₁₀₀ / 1.36 ₁₀₀	0.27 ₁₀₀ / 0.28 ₁₀₀ / 1.39 ₁₀₀	n.a. / n.a. / n.a.	0.60 ₁₀₀ / 0.71 ₁₀₀ / 0.76 ₁₀₀	0.60 ₁₀₀ / 0.78 ₁₀₀ / 0.97 ₁₀₀	0.48 ₁₀₀ / 0.45 ₁₀₀ / 0.58 ₁₀₀	0.48 ₁₀₀ / 0.53 ₁₀₀ / 0.67 ₁₀₀
	GCN	0.15 ₁₀₀ / 0.16 ₁₀₀ / 1.47 ₁₀₀	0.18 ₁₀₀ / 0.18 ₁₀₀ / 1.42 ₁₀₀	n.a. / n.a. / n.a.	0.34 ₁₀₀ / 0.32 ₁₀₀ / 0.35 ₁₀₀	0.34 ₁₀₀ / 0.31 ₁₀₀ / 0.34 ₁₀₀	0.28 ₁₀₀ / 0.27 ₁₀₀ / 0.30 ₁₀₀	0.13 ₁₀₀ / 0.21 ₁₀₀ / 0.61 ₁₀₀
	GCN (Ours)	0.14 ₁₀₀ / 0.14 ₁₀₀ / 1.51 ₁₀₀	0.15 ₁₀₀ / 0.16 ₁₀₀ / 1.48 ₁₀₀	n.a. / n.a. / n.a.	0.32 ₁₀₀ / 0.36 ₁₀₀ / 0.47 ₁₀₀	0.32 ₁₀₀ / 0.76 ₁₀₀ / 1.09 ₁₀₀	0.23 ₁₀₀ / 0.22 ₁₀₀ / 0.27 ₁₀₀	0.14 ₁₀₀ / 0.10 ₁₀₀ / 0.40 ₁₀₀
	BGCN	0.23 ₁₀₀ / 0.23 ₁₀₀ / 1.54 ₁₀₀	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.55 ₁₀₀	0.36 ₁₀₀ / 0.41 ₁₀₀ / 0.61 ₁₀₀	0.36 ₁₀₀ / 0.53 ₁₀₀ / 0.87 ₁₀₀	0.36 ₁₀₀ / 0.98 ₁₀₀ / 1.32 ₁₀₀	0.33 ₁₀₀ / 0.30 ₁₀₀ / 0.24 ₁₀₀	0.33 ₁₀₀ / 0.22 ₁₀₀ / 0.29 ₁₀₀
	GCN-MCD	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.43 ₁₀₀	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.43 ₁₀₀	0.34 ₁₀₀ / 0.41 ₁₀₀ / 0.60 ₁₀₀	0.34 ₁₀₀ / 0.46 ₁₀₀ / 0.71 ₁₀₀	0.34 ₁₀₀ / 0.97 ₁₀₀ / 1.39 ₁₀₀	0.31 ₁₀₀ / 0.28 ₁₀₀ / 0.26 ₁₀₀	0.32 ₁₀₀ / 0.24 ₁₀₀ / 0.29 ₁₀₀
Published	GCN-DE	0.28 ₁₀₀ / 0.28 ₁₀₀ / 1.40 ₁₀₀	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.40 ₁₀₀	0.40 ₁₀₀ / 0.47 ₁₀₀ / 0.71 ₁₀₀	0.40 ₁₀₀ / 0.50 ₁₀₀ / 0.75 ₁₀₀	0.40 ₁₀₀ / 0.87 ₁₀₀ / 1.30 ₁₀₀	0.36 ₁₀₀ / 0.34 ₁₀₀ / 0.34 ₁₀₀	0.37 ₁₀₀ / 0.32 ₁₀₀ / 0.37 ₁₀₀
	GCN-Ens	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.42 ₁₀₀	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.43 ₁₀₀	0.34 ₁₀₀ / 0.41 ₁₀₀ / 0.60 ₁₀₀	0.34 ₁₀₀ / 0.43 ₁₀₀ / 0.65 ₁₀₀	0.34 ₁₀₀ / 0.95 ₁₀₀ / 1.32 ₁₀₀	0.31 ₁₀₀ / 0.28 ₁₀₀ / 0.25 ₁₀₀	0.32 ₁₀₀ / 0.24 ₁₀₀ / 0.29 ₁₀₀
	GCN-MCD+DE	0.28 ₁₀₀ / 0.28 ₁₀₀ / 1.40 ₁₀₀	0.21 ₁₀₀ / 0.21 ₁₀₀ / 1.40 ₁₀₀	0.40 ₁₀₀ / 0.48 ₁₀₀ / 0.71 ₁₀₀	0.40 ₁₀₀ / 0.49 ₁₀₀ / 0.74 ₁₀₀	0.40 ₁₀₀ / 0.86 ₁₀₀ / 1.30 ₁₀₀	0.36 ₁₀₀ / 0.34 ₁₀₀ / 0.34 ₁₀₀	0.36 ₁₀₀ / 0.32 ₁₀₀ / 0.36 ₁₀₀
	SGCN	0.33 ₁₀₀ / 0.33 ₁₀₀ / 1.28 ₁₀₀	0.30 ₁₀₀ / 0.30 ₁₀₀ / 1.28 ₁₀₀	0.53 ₁₀₀ / 0.65 ₁₀₀ / 0.74 ₁₀₀	0.53 ₁₀₀ / 0.65 ₁₀₀ / 0.74 ₁₀₀	0.53 ₁₀₀ / 0.89 ₁₀₀ / 1.17 ₁₀₀	0.47 ₁₀₀ / 0.45 ₁₀₀ / 0.54 ₁₀₀	0.51 ₁₀₀ / 0.52 ₁₀₀ / 0.55 ₁₀₀
	GCN	0.21 ₁₀₀ / 0.21 ₁₀₀ / 1.46 ₁₀₀	0.13 ₁₀₀ / 0.13 ₁₀₀ / 1.50 ₁₀₀	0.33 ₁₀₀ / 0.32 ₁₀₀ / 0.43 ₁₀₀	0.33 ₁₀₀ / 0.33 ₁₀₀ / 0.47 ₁₀₀	0.33 ₁₀₀ / 0.27 ₁₀₀ / 0.32 ₁₀₀	0.31 ₁₀₀ / 0.28 ₁₀₀ / 0.30 ₁₀₀	0.31 ₁₀₀ / 0.29 ₁₀₀ / 0.28 ₁₀₀
	GCN (Ours)	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.42 ₁₀₀	0.15 ₁₀₀ / 0.15 ₁₀₀ / 1.43 ₁₀₀	0.34 ₁₀₀ / 0.41 ₁₀₀ / 0.60 ₁₀₀	0.34 ₁₀₀ / 0.44 ₁₀₀ / 0.69 ₁₀₀	0.34 ₁₀₀ / 0.99 ₁₀₀ / 1.38 ₁₀₀	0.31 ₁₀₀ / 0.28 ₁₀₀ / 0.25 ₁₀₀	0.32 ₁₀₀ / 0.24 ₁₀₀ / 0.30 ₁₀₀
Amazon Photo	BGCN	0.13 ₁₀₀ / 0.13 ₁₀₀ / 1.91 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 1.76 ₁₀₀	n.a. / n.a. / n.a.	0.33 ₁₀₀ / 0.40 ₁₀₀ / 0.57 ₁₀₀	0.33 ₁₀₀ / 0.76 ₁₀₀ / 1.06 ₁₀₀	0.36 ₁₀₀ / 0.37 ₁₀₀ / 0.34 ₁₀₀	0.27 ₁₀₀ / 0.19 ₁₀₀ / 0.50 ₁₀₀
	GCN-MCD	0.12 ₁₀₀ / 0.12 ₁₀₀ / 1.76 ₁₀₀	0.26 ₁₀₀ / 0.26 ₁₀₀ / 1.63 ₁₀₀	n.a. / n.a. / n.a.	0.33 ₁₀₀ / 0.39 ₁₀₀ / 0.52 ₁₀₀	0.33 ₁₀₀ / 0.72 ₁₀₀ / 1.04 ₁₀₀	0.31 ₁₀₀ / 0.31 ₁₀₀ / 0.31 ₁₀₀	0.25 ₁₀₀ / 0.20 ₁₀₀ / 0.43 ₁₀₀
	GCN-DE	0.12 ₁₀₀ / 0.12 ₁₀₀ / 1.76 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 1.62 ₁₀₀	n.a. / n.a. / n.a.	0.36 ₁₀₀ / 0.41 ₁₀₀ / 0.59 ₁₀₀	0.36 ₁₀₀ / 0.71 ₁₀₀ / 1.07 ₁₀₀	0.33 ₁₀₀ / 0.33 ₁₀₀ / 0.34 ₁₀₀	0.28 ₁₀₀ / 0.25 ₁₀₀ / 0.43 ₁₀₀
	GCN-Ens	0.12 ₁₀₀ / 0.12 ₁₀₀ / 1.77 ₁₀₀	0.26 ₁₀₀ / 0.26 ₁₀₀ / 1.63 ₁₀₀	n.a. / n.a. / n.a.	0.33 ₁₀₀ / 0.38 ₁₀₀ / 0.52 ₁₀₀	0.33 ₁₀₀ / 0.71 ₁₀₀ / 1.02 ₁₀₀	0.30 ₁₀₀ / 0.31 ₁₀₀ / 0.31 ₁₀₀	0.25 ₁₀₀ / 0.20 ₁₀₀ / 0.43 ₁₀₀
	GCN-MCD+DE	0.12 ₁₀₀ / 0.12 ₁₀₀ / 1.76 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 1.62 ₁₀₀	n.a. / n.a. / n.a.	0.36 ₁₀₀ / 0.41 ₁₀₀ / 0.60 ₁₀₀	0.36 ₁₀₀ / 0.70 ₁₀₀ / 1.06 ₁₀₀	0.33 ₁₀₀ / 0.33 ₁₀₀ / 0.34 ₁₀₀	0.28 ₁₀₀ / 0.25 ₁₀₀ / 0.44 ₁₀₀
	SGCN	0.13 ₁₀₀ / 0.14 ₁₀₀ / 1.66 ₁₀₀	0.26 ₁₀₀ / 0.26 ₁₀₀ / 1.59 ₁₀₀	n.a. / n.a. / n.a.	0.37 ₁₀₀ / 0.45 ₁₀₀ / 0.54 ₁₀₀	0.37 ₁₀₀ / 0.65 ₁₀₀ / 0.91 ₁₀₀	0.33 ₁₀₀ / 0.32 ₁₀₀ / 0.37 ₁₀₀	0.30 ₁₀₀ / 0.33 ₁₀₀ / 0.45 ₁₀₀
Amazon Computers	GCN	0.11 ₁₀₀ / 0.12 ₁₀₀ / 1.75 ₁₀₀	0.22 ₁₀₀ / 0.22 ₁₀₀ / 1.67 ₁₀₀	n.a. / n.a. / n.a.	0.33 ₁₀₀ / 0.36 ₁₀₀ / 0.46 ₁₀₀	0.33 ₁₀₀ / 0.31 ₁₀₀ / 0.36 ₁₀₀	0.29 ₁₀₀ / 0.29 ₁₀₀ / 0.28 ₁₀₀	0.21 ₁₀₀ / 0.22 ₁₀₀ / 0.52 ₁₀₀
	GCN (Ours)	0.13 ₁₀₀ / 0.13 ₁₀₀ / 1.77 ₁₀₀	0.26 ₁₀₀ / 0.26 ₁₀₀ / 1.63 ₁₀₀	n.a. / n.a. / n.a.	0.38 ₁₀₀ / 0.39 ₁₀₀ / 0.52 ₁₀₀	0.38 ₁₀₀ / 0.72 ₁₀₀ / 1.05 ₁₀₀	0.31 ₁₀₀ / 0.31 ₁₀₀ / 0.31 ₁₀₀	0.25 ₁₀₀ / 0.20 ₁₀₀ / 0.43 ₁₀₀
	BGCN	0.13 ₁₀₀ / 0.13 ₁₀₀ / 1.81 ₁₀₀	0.07 ₁₀₀ / 0.07 ₁₀₀ / 1.73 ₁₀₀	0.17 ₁₀₀ / 0.27 ₁₀₀ / 0.34 ₁₀₀	0.17 ₁₀₀ / 0.65 ₁₀₀ / 0.76 ₁₀₀	0.17 ₁₀₀ / 0.89 ₁₀₀ / 1.02 ₁₀₀	0.15 ₁₀₀ / 0.13 ₁₀₀ / 0.18 ₁₀₀	0.04 ₁₀₀ / 0.05 ₁₀₀ / 0.27 ₁₀₀
	GCN-MCD	0.12 ₁₀₀ / 0.12 ₁₀₀ / 1.62 ₁₀₀	0.09 ₁₀₀ / 0.09 ₁₀₀ / 1.59 ₁₀₀	0.17 ₁₀₀ / 0.28 ₁₀₀ / 0.35 ₁₀₀	0.17 ₁₀₀ / 0.48 ₁₀₀ / 0.57 ₁₀₀	0.17 ₁₀₀ / 0.86 ₁₀₀ / 1.02 ₁₀₀	0.13 ₁₀₀ / 0.12 ₁₀₀ / 0.23 ₁₀₀	0.05 ₁₀₀ / 0.08 ₁₀₀ / 0.20 ₁₀₀
	GCN-DE	0.13 ₁₀₀ / 0.14 ₁₀₀ / 1.57 ₁₀₀	0.10 ₁₀₀ / 0.10 ₁₀₀ / 1.55 ₁₀₀	0.19 ₁₀₀ / 0.32 ₁₀₀ / 0.40 ₁₀₀	0.19 ₁₀₀ / 0.46 ₁₀₀ / 0.55 ₁₀₀	0.19 ₁₀₀ / 0.85 ₁₀₀ / 1.00 ₁₀₀	0.15 ₁₀₀ / 0.13 ₁₀₀ / 0.25 ₁₀₀	0.07 ₁₀₀ / 0.10 ₁₀₀ / 0.29 ₁₀₀
	GCN-Ens	0.12 ₁₀₀ / 0.13 ₁₀₀ / 1.62 ₁₀₀	0.08 ₁₀₀ / 0.09 ₁₀₀ / 1.59 ₁₀₀	0.17 ₁₀₀ / 0.28 ₁₀₀ / 0.34 ₁₀₀	0.17 ₁₀₀ / 0.47 ₁₀₀ / 0.56 ₁₀₀	0.17 ₁₀₀ / 0.83 ₁₀₀ / 0.98 ₁₀₀	0.13 ₁₀₀ / 0.12 ₁₀₀ / 0.25 ₁₀₀	0.05 ₁₀₀ / 0.08 ₁₀₀ / 0.26 ₁₀₀

	Model	LoC (<i>last</i>)	LoC (<i>hetero</i>)	Ber(0.5)	Ber($\hat{p}$) (<i>near</i>)	$\mathcal{N}(0, I)$ (<i>far</i>)	Homophily	Page Rank
GCN	EBM	85.5 $\pm$ 1.8	75.1 $\pm$ 5.5	53.9 $\pm$ 3.9	63.5 $\pm$ 2.3	36.1 $\pm$ 2.1	66.3 $\pm$ 2.4	71.6 $\pm$ 3.5
	Safe	87.3 $\pm$ 2.2	76.5 $\pm$ 5.5	51.0 $\pm$ 2.4	55.3 $\pm$ 1.9	40.3 $\pm$ 2.7	63.7 $\pm$ 2.1	54.0 $\pm$ 0.6
	GEBM	88.0 $\pm$ 1.5	75.1 $\pm$ 3.3	94.3 $\pm$ 2.5	70.6 $\pm$ 2.5	88.9 $\pm$ 3.4	68.8 $\pm$ 2.2	56.8 $\pm$ 1.1
GAT	EBM	86.4 $\pm$ 2.9	80.3 $\pm$ 2.9	56.3 $\pm$ 3.8	56.5 $\pm$ 1.5	45.2 $\pm$ 2.5	67.0 $\pm$ 1.3	55.1 $\pm$ 0.8
	Safe	88.9 $\pm$ 1.4	83.1 $\pm$ 3.0	53.5 $\pm$ 2.9	54.2 $\pm$ 1.2	45.4 $\pm$ 2.3	62.1 $\pm$ 1.7	53.3 $\pm$ 1.0
	GEBM	79.5 $\pm$ 8.4	64.4 $\pm$ 12.0	88.6 $\pm$ 6.8	61.9 $\pm$ 4.7	81.8 $\pm$ 4.1	66.5 $\pm$ 3.0	54.7 $\pm$ 1.3
GIN	EBM	70.2 $\pm$ 8.2	49.1 $\pm$ 9.9	46.0 $\pm$ 1.9	52.8 $\pm$ 2.6	44.0 $\pm$ 1.9	51.4 $\pm$ 3.5	65.7 $\pm$ 1.0
	Safe	72.3 $\pm$ 6.4	51.4 $\pm$ 7.1	47.9 $\pm$ 1.1	50.3 $\pm$ 0.9	46.6 $\pm$ 2.1	46.1 $\pm$ 2.1	50.1 $\pm$ 0.4
	GEBM	75.1 $\pm$ 5.9	52.7 $\pm$ 8.2	51.2 $\pm$ 2.6	51.4 $\pm$ 2.6	52.9 $\pm$ 2.5	54.7 $\pm$ 2.6	54.0 $\pm$ 2.0
SAGE	EBM	66.6 $\pm$ 8.2	55.3 $\pm$ 13.4	46.4 $\pm$ 2.4	53.5 $\pm$ 3.7	43.4 $\pm$ 2.8	52.5 $\pm$ 4.3	65.4 $\pm$ 2.0
	Safe	70.2 $\pm$ 6.8	56.4 $\pm$ 10.6	48.5 $\pm$ 1.1	49.9 $\pm$ 1.3	46.9 $\pm$ 1.8	46.5 $\pm$ 2.7	50.3 $\pm$ 0.8
	GEBM	70.6 $\pm$ 10.4	49.3 $\pm$ 10.2	51.5 $\pm$ 2.4	51.4 $\pm$ 2.4	53.1 $\pm$ 2.7	55.1 $\pm$ 3.3	54.5 $\pm$ 2.5

Table 25: O.o.d. detection AUC-PR($\uparrow$) using different backbones on CoraML in an inductive setting.

	Model	LoC (<i>last</i>)	LoC (<i>hetero.</i>)	Ber($\hat{p}$) (<i>near</i>)	Ber(0.5)	$\mathcal{N}(0, 1)$ (<i>far</i>)	Page Rank	Homophily	Rank($\downarrow$)
CoraML	EBM	89.9 $\pm$ 0.9	88.4 $\pm$ 2.3	67.3 $\pm$ 1.9	58.6 $\pm$ 4.6	25.7 $\pm$ 1.3	74.2 $\pm$ 2.1	71.2 $\pm$ 1.6	3.9
	Indep.	81.8 $\pm$ 1.6	78.5 $\pm$ 1.4	87.1 $\pm$ 2.0	99.5 $\pm$ 0.7	96.3 $\pm$ 3.3	63.9 $\pm$ 1.7	69.0 $\pm$ 1.7	3.7
	Local	87.5 $\pm$ 2.0	86.9 $\pm$ 2.1	57.4 $\pm$ 1.6	57.9 $\pm$ 2.6	63.1 $\pm$ 2.9	51.0 $\pm$ 0.5	81.3 $\pm$ 0.8	4.3
	Group	92.8 $\pm$ 0.6	90.2 $\pm$ 1.1	55.5 $\pm$ 1.0	60.9 $\pm$ 1.9	63.7 $\pm$ 2.5	49.5 $\pm$ 1.1	71.8 $\pm$ 1.7	3.4
	GEBM- E_θ	92.2 $\pm$ 0.6	89.9 $\pm$ 2.0	75.9 $\pm$ 1.6	72.3 $\pm$ 5.7	1.2 $\pm$ 0.4	57.6 $\pm$ 1.4	74.3 $\pm$ 1.6	3.2
	GEBM	91.6 $\pm$ 0.8	89.5 $\pm$ 1.7	77.7 $\pm$ 2.0	94.5 $\pm$ 2.4	86.4 $\pm$ 3.6	57.7 $\pm$ 1.3	76.7 $\pm$ 1.3	2.5
CoraML-LLM	EBM	87.1 $\pm$ 0.9	87.5 $\pm$ 3.3	n.a.	70.1 $\pm$ 1.6	33.0 $\pm$ 2.5	72.7 $\pm$ 1.4	69.7 $\pm$ 0.7	3.3
	Indep.	78.9 $\pm$ 1.1	82.0 $\pm$ 2.7	n.a.	92.1 $\pm$ 2.9	96.0 $\pm$ 2.0	53.2 $\pm$ 0.7	65.1 $\pm$ 1.0	3.6
	Local	84.5 $\pm$ 1.2	84.9 $\pm$ 2.4	n.a.	55.9 $\pm$ 1.2	62.4 $\pm$ 2.8	50.9 $\pm$ 0.4	80.6 $\pm$ 1.1	3.7
	Group	91.6 $\pm$ 1.2	91.4 $\pm$ 1.3	n.a.	58.5 $\pm$ 2.4	62.4 $\pm$ 3.4	49.6 $\pm$ 0.7	70.7 $\pm$ 1.5	2.9
	GEBM- E_θ	89.8 $\pm$ 1.0	90.4 $\pm$ 1.7	n.a.	88.5 $\pm$ 1.3	7.1 $\pm$ 1.4	51.2 $\pm$ 0.6	73.1 $\pm$ 0.9	2.7
	GEBM	88.4 $\pm$ 1.1	89.7 $\pm$ 2.0	n.a.	82.2 $\pm$ 3.8	85.3 $\pm$ 3.4	51.9 $\pm$ 0.7	75.0 $\pm$ 1.2	2.4
Citeseer	EBM	87.4 $\pm$ 2.5	88.0 $\pm$ 1.9	64.7 $\pm$ 1.5	43.8 $\pm$ 6.5	23.5 $\pm$ 1.6	66.4 $\pm$ 2.0	52.0 $\pm$ 2.1	3.9
	Indep.	78.2 $\pm$ 2.7	76.5 $\pm$ 1.4	69.3 $\pm$ 1.1	99.6 $\pm$ 0.8	75.5 $\pm$ 8.6	50.6 $\pm$ 1.0	51.9 $\pm$ 1.4	3.7
	Local	86.9 $\pm$ 2.2	90.8 $\pm$ 1.3	60.0 $\pm$ 1.3	68.6 $\pm$ 4.5	63.0 $\pm$ 2.0	47.6 $\pm$ 2.1	55.5 $\pm$ 1.9	3.3
	Group	90.6 $\pm$ 1.6	88.8 $\pm$ 1.1	55.5 $\pm$ 1.3	70.4 $\pm$ 2.8	61.7 $\pm$ 3.0	46.2 $\pm$ 2.0	53.0 $\pm$ 2.2	3.6
	GEBM- E_θ	90.0 $\pm$ 2.2	87.8 $\pm$ 1.7	63.9 $\pm$ 1.0	45.4 $\pm$ 10.4	5.6 $\pm$ 1.0	48.2 $\pm$ 1.9	53.2 $\pm$ 2.2	4.0
	GEBM	88.9 $\pm$ 2.1	88.7 $\pm$ 1.4	66.0 $\pm$ 1.2	96.3 $\pm$ 2.9	72.1 $\pm$ 3.9	48.3 $\pm$ 1.8	53.7 $\pm$ 2.0	2.5
PubMed	EBM	65.8 $\pm$ 1.7	67.8 $\pm$ 2.4	n.a.	48.6 $\pm$ 4.5	27.5 $\pm$ 1.1	72.6 $\pm$ 2.3	52.2 $\pm$ 1.5	3.9
	Indep.	66.0 $\pm$ 0.9	59.6 $\pm$ 3.4	n.a.	88.6 $\pm$ 7.9	82.5 $\pm$ 8.3	50.9 $\pm$ 1.6	54.8 $\pm$ 1.0	3.2
	Local	68.3 $\pm$ 1.0	66.3 $\pm$ 4.9	n.a.	56.1 $\pm$ 3.2	61.4 $\pm$ 1.2	47.7 $\pm$ 1.2	62.6 $\pm$ 1.6	3.4
	Group	72.8 $\pm$ 2.2	71.4 $\pm$ 4.9	n.a.	59.5 $\pm$ 2.3	60.3 $\pm$ 1.4	50.6 $\pm$ 2.0	56.4 $\pm$ 2.0	2.3
	GEBM- E_θ	68.5 $\pm$ 2.1	69.2 $\pm$ 2.6	n.a.	52.6 $\pm$ 10.1	8.1 $\pm$ 0.7	49.7 $\pm$ 1.5	58.9 $\pm$ 1.1	3.2
	GEBM	70.9 $\pm$ 1.0	67.2 $\pm$ 5.1	n.a.	84.0 $\pm$ 7.4	75.7 $\pm$ 4.2	49.4 $\pm$ 1.6	59.0 $\pm$ 1.7	2.6
Amazon Photo	EBM	75.2 $\pm$ 5.0	78.1 $\pm$ 3.2	57.8 $\pm$ 0.6	52.1 $\pm$ 1.7	42.6 $\pm$ 1.3	90.3 $\pm$ 0.9	60.0 $\pm$ 0.8	4.1
	Indep.	75.6 $\pm$ 3.7	76.3 $\pm$ 3.2	91.4 $\pm$ 2.2	68.6 $\pm$ 10.6	95.8 $\pm$ 2.9	59.7 $\pm$ 1.2	61.4 $\pm$ 1.7	3.4
	Local	73.1 $\pm$ 4.1	90.6 $\pm$ 1.4	51.4 $\pm$ 1.1	49.2 $\pm$ 1.0	57.3 $\pm$ 1.3	54.0 $\pm$ 0.6	68.4 $\pm$ 0.8	3.9
	Group	89.9 $\pm$ 5.4	82.3 $\pm$ 4.1	52.8 $\pm$ 1.1	50.7 $\pm$ 1.5	57.6 $\pm$ 1.2	53.5 $\pm$ 0.8	59.6 $\pm$ 2.1	4.3
	GEBM- E_θ	85.4 $\pm$ 4.2	85.5 $\pm$ 3.3	92.5 $\pm$ 1.8	71.7 $\pm$ 8.6	0.5 $\pm$ 0.2	62.0 $\pm$ 0.6	65.9 $\pm$ 1.1	2.6
	GEBM	83.6 $\pm$ 5.2	87.5 $\pm$ 2.9	84.8 $\pm$ 3.1	64.2 $\pm$ 8.6	92.4 $\pm$ 2.8	58.0 $\pm$ 0.7	68.3 $\pm$ 1.1	2.7
Amazon Computers	EBM	75.9 $\pm$ 5.1	85.5 $\pm$ 3.6	56.6 $\pm$ 0.7	54.9 $\pm$ 2.1	40.1 $\pm$ 1.1	85.3 $\pm$ 1.4	61.5 $\pm$ 0.9	3.6
	Indep.	74.4 $\pm$ 3.8	79.2 $\pm$ 2.9	90.0 $\pm$ 1.9	82.1 $\pm$ 6.8	96.1 $\pm$ 7.0	51.8 $\pm$ 1.9	59.2 $\pm$ 0.9	4.1
	Local	75.7 $\pm$ 3.3	85.1 $\pm$ 6.1	51.8 $\pm$ 0.6	50.9 $\pm$ 1.1	57.4 $\pm$ 1.4	54.0 $\pm$ 0.4	69.2 $\pm$ 1.1	3.9
	Group	86.6 $\pm$ 4.9	89.2 $\pm$ 4.0	53.3 $\pm$ 0.6	52.3 $\pm$ 1.0	58.8 $\pm$ 1.3	51.3 $\pm$ 1.6	56.1 $\pm$ 2.6	4.1
	GEBM- E_θ	86.7 $\pm$ 3.6	89.7 $\pm$ 2.3	89.9 $\pm$ 1.3	84.3 $\pm$ 5.9	0.1 $\pm$ 0.1	53.0 $\pm$ 1.4	65.3 $\pm$ 0.8	2.2
	GEBM	81.8 $\pm$ 3.5	87.9 $\pm$ 3.9	85.2 $\pm$ 2.2	75.4 $\pm$ 6.2	94.1 $\pm$ 6.6	52.6 $\pm$ 1.3	64.4 $\pm$ 1.5	3.1
Coauthor-CS	EBM	89.9 $\pm$ 1.6	94.2 $\pm$ 1.1	69.6 $\pm$ 0.8	71.6 $\pm$ 1.6	28.8 $\pm$ 0.7	81.9 $\pm$ 0.8	60.5 $\pm$ 0.7	3.8
	Indep.	90.3 $\pm$ 0.9	92.1 $\pm$ 1.8	99.2 $\pm$ 0.3	100.0 $\pm$ 0.0	92.7 $\pm$ 5.4	64.4 $\pm$ 0.9	58.2 $\pm$ 0.5	3.2
	Local	73.8 $\pm$ 2.5	87.6 $\pm$ 2.1	57.0 $\pm$ 1.1	59.1 $\pm$ 1.0	62.7 $\pm$ 2.3	45.4 $\pm$ 0.3	77.0 $\pm$ 0.9	4.9
	Group	96.0 $\pm$ 1.2	95.6 $\pm$ 1.4	62.3 $\pm$ 1.8	65.3 $\pm$ 1.5	63.9 $\pm$ 2.0	51.4 $\pm$ 0.6	58.5 $\pm$ 1.7	3.8
	GEBM- E_θ	94.5 $\pm$ 1.4	97.5 $\pm$ 0.6	95.5 $\pm$ 0.4	97.5 $\pm$ 0.7	0.2 $\pm$ 0.1	59.9 $\pm$ 0.8	64.7 $\pm$ 0.7	2.8
	GEBM	92.8 $\pm$ 1.5	96.1 $\pm$ 1.3	95.6 $\pm$ 1.1	99.5 $\pm$ 0.2	90.0 $\pm$ 5.4	56.5 $\pm$ 0.9	69.5 $\pm$ 1.0	2.5
Coauthor-Physics	EBM	93.4 $\pm$ 1.8	97.0 $\pm$ 0.6	59.4 $\pm$ 1.3	62.4 $\pm$ 1.4	39.5 $\pm$ 1.0	82.0 $\pm$ 1.6	56.5 $\pm$ 0.6	3.6
	Indep.	89.6 $\pm$ 2.6	94.6 $\pm$ 1.0	91.9 $\pm$ 1.5	99.9 $\pm$ 0.1	73.4 $\pm$ 7.4	65.7 $\pm$ 1.2	55.7 $\pm$ 0.4	3.7
	Local	91.7 $\pm$ 1.9	96.3 $\pm$ 0.6	53.4 $\pm$ 0.9	55.9 $\pm$ 0.9	58.0 $\pm$ 1.1	52.3 $\pm$ 0.5	66.1 $\pm$ 1.2	4.5
	Group	95.5 $\pm$ 1.9	98.3 $\pm$ 0.3	54.7 $\pm$ 1.0	57.0 $\pm$ 1.0	56.7 $\pm$ 1.3	54.7 $\pm$ 0.7	56.4 $\pm$ 1.0	4.1
	GEBM- E_θ	96.6 $\pm$ 0.9	98.4 $\pm$ 0.4	81.2 $\pm$ 2.9	90.0 $\pm$ 1.7	3.6 $\pm$ 0.6	62.2 $\pm$ 1.3	58.3 $\pm$ 0.6	2.8
	GEBM	94.9 $\pm$ 1.8	98.4 $\pm$ 0.3	81.8 $\pm$ 2.5	95.3 $\pm$ 1.5	72.7 $\pm$ 5.8	60.1 $\pm$ 1.1	61.3 $\pm$ 0.8	2.4

Table 26: O.o.d-detection AUC-ROC ($\uparrow$) using different EBMs in an inductive setting.

	Model	LoC (<i>last</i>)	LoC (<i>hetero.</i>)	Ber($\hat{p}$) (<i>near</i>)	Ber(0.5)	$\mathcal{N}(0, 1)$ (<i>far</i>)	Page Rank	Homophily	Rank()
CoraML	EBM	85.5 $\pm$ 1.8	75.1 $\pm$ 5.5	63.8 $\pm$ 3.7	53.9 $\pm$ 3.9	35.9 $\pm$ 1.8	71.6 $\pm$ 3.5	66.3 $\pm$ 2.4	3.2
	Indep.	74.7 $\pm$ 2.0	58.1 $\pm$ 2.3	82.3 $\pm$ 2.4	99.5 $\pm$ 0.7	97.9 $\pm$ 1.8	62.9 $\pm$ 1.8	61.4 $\pm$ 2.1	3.6
	Local	78.2 $\pm$ 4.3	68.8 $\pm$ 4.4	56.0 $\pm$ 2.7	57.1 $\pm$ 2.4	62.5 $\pm$ 3.3	54.9 $\pm$ 0.4	77.7 $\pm$ 1.1	4.1
	Group	89.7 $\pm$ 1.2	73.1 $\pm$ 2.1	54.7 $\pm$ 0.8	60.9 $\pm$ 2.0	61.7 $\pm$ 2.4	51.1 $\pm$ 0.9	61.4 $\pm$ 2.2	4.3
	GEBM- E_θ	88.8 $\pm$ 1.2	74.3 $\pm$ 4.7	69.2 $\pm$ 2.4	63.4 $\pm$ 6.3	30.6 $\pm$ 1.5	55.9 $\pm$ 1.1	64.5 $\pm$ 2.5	3.5
	GEBM	88.0 $\pm$ 1.5	75.1 $\pm$ 3.3	71.3 $\pm$ 2.9	94.6 $\pm$ 2.5	88.8 $\pm$ 3.1	56.8 $\pm$ 1.1	68.8 $\pm$ 2.2	2.2
CoraML-LLM	EBM	84.0 $\pm$ 2.3	75.9 $\pm$ 6.5	n.a.	65.1 $\pm$ 2.7	39.2 $\pm$ 1.5	69.1 $\pm$ 1.8	66.5 $\pm$ 1.0	3.0
	Indep.	69.6 $\pm$ 2.4	63.0 $\pm$ 5.1	n.a.	82.1 $\pm$ 5.4	97.5 $\pm$ 1.2	54.7 $\pm$ 0.8	58.6 $\pm$ 1.3	3.5
	Local	75.2 $\pm$ 3.7	61.9 $\pm$ 4.9	n.a.	52.1 $\pm$ 1.3	60.5 $\pm$ 3.9	54.4 $\pm$ 0.4	77.6 $\pm$ 1.5	3.5
	Group	88.1 $\pm$ 1.4	77.9 $\pm$ 4.1	n.a.	56.3 $\pm$ 2.0	59.5 $\pm$ 3.4	51.6 $\pm$ 0.6	59.7 $\pm$ 1.8	3.3
	GEBM- E_θ	87.7 $\pm$ 1.2	78.8 $\pm$ 3.9	n.a.	83.2 $\pm$ 2.3	30.7 $\pm$ 0.9	52.4 $\pm$ 0.5	64.5 $\pm$ 1.2	2.8
	GEBM	84.7 $\pm$ 2.0	76.8 $\pm$ 5.1	n.a.	72.8 $\pm$ 4.9	87.3 $\pm$ 3.8	53.2 $\pm$ 0.6	67.5 $\pm$ 2.1	2.6
Citeseer	EBM	71.0 $\pm$ 3.8	76.3 $\pm$ 4.3	59.0 $\pm$ 2.4	42.5 $\pm$ 3.3	34.8 $\pm$ 1.4	59.1 $\pm$ 2.1	51.8 $\pm$ 2.1	3.6
	Indep.	51.8 $\pm$ 3.2	59.4 $\pm$ 1.6	63.5 $\pm$ 1.7	99.6 $\pm$ 0.8	85.4 $\pm$ 5.5	49.8 $\pm$ 0.8	51.3 $\pm$ 1.1	3.7
	Local	63.7 $\pm$ 6.5	82.0 $\pm$ 2.8	57.5 $\pm$ 2.3	67.2 $\pm$ 6.0	62.5 $\pm$ 2.6	48.5 $\pm$ 1.4	55.5 $\pm$ 1.9	2.9
	Group	71.3 $\pm$ 3.6	75.7 $\pm$ 2.6	52.7 $\pm$ 1.7	70.9 $\pm$ 3.9	60.5 $\pm$ 3.4	45.7 $\pm$ 1.4	52.1 $\pm$ 1.8	4.0
	GEBM- E_θ	74.3 $\pm$ 3.9	74.1 $\pm$ 3.8	55.3 $\pm$ 1.9	42.9 $\pm$ 5.0	30.7 $\pm$ 1.4	46.8 $\pm$ 1.4	52.1 $\pm$ 1.7	4.2
	GEBM	69.4 $\pm$ 4.5	78.5 $\pm$ 3.0	59.3 $\pm$ 2.0	96.6 $\pm$ 2.7	75.3 $\pm$ 3.9	47.2 $\pm$ 1.4	52.4 $\pm$ 1.8	2.7
PubMed	EBM	53.5 $\pm$ 2.1	33.0 $\pm$ 3.3	n.a.	46.1 $\pm$ 3.2	39.1 $\pm$ 0.7	66.0 $\pm$ 3.2	53.7 $\pm$ 1.7	3.4
	Indep.	51.7 $\pm$ 1.5	24.8 $\pm$ 2.7	n.a.	89.1 $\pm$ 9.0	89.5 $\pm$ 5.0	50.5 $\pm$ 1.3	53.3 $\pm$ 1.2	3.2
	Local	50.4 $\pm$ 1.3	29.0 $\pm$ 4.5	n.a.	55.8 $\pm$ 3.4	61.1 $\pm$ 1.5	47.5 $\pm$ 0.7	61.4 $\pm$ 1.7	3.8
	Group	60.6 $\pm$ 3.0	40.6 $\pm$ 7.7	n.a.	59.1 $\pm$ 2.9	59.1 $\pm$ 1.2	50.4 $\pm$ 1.7	52.3 $\pm$ 1.7	2.6
	GEBM- E_θ	57.2 $\pm$ 3.0	39.0 $\pm$ 3.7	n.a.	47.7 $\pm$ 5.9	31.8 $\pm$ 0.4	49.0 $\pm$ 1.2	56.0 $\pm$ 1.3	3.1
	GEBM	56.7 $\pm$ 1.7	33.4 $\pm$ 6.1	n.a.	83.8 $\pm$ 8.8	79.8 $\pm$ 3.2	49.0 $\pm$ 1.2	56.1 $\pm$ 1.8	2.5
Amazon Photo	EBM	63.3 $\pm$ 7.7	51.4 $\pm$ 4.2	59.0 $\pm$ 0.7	52.0 $\pm$ 1.7	45.1 $\pm$ 0.7	89.0 $\pm$ 1.4	62.1 $\pm$ 1.3	3.8
	Indep.	65.6 $\pm$ 5.3	48.9 $\pm$ 4.5	88.7 $\pm$ 2.8	58.9 $\pm$ 9.5	97.6 $\pm$ 1.6	57.2 $\pm$ 1.3	60.3 $\pm$ 2.2	3.4
	Local	55.6 $\pm$ 6.7	73.9 $\pm$ 4.5	50.8 $\pm$ 1.6	49.1 $\pm$ 1.0	56.9 $\pm$ 1.0	55.0 $\pm$ 0.7	65.8 $\pm$ 1.2	4.0
	Group	80.6 $\pm$ 7.6	49.4 $\pm$ 5.3	52.5 $\pm$ 1.3	49.8 $\pm$ 1.2	56.7 $\pm$ 1.1	51.2 $\pm$ 0.6	55.0 $\pm$ 1.8	4.6
	GEBM- E_θ	76.2 $\pm$ 7.1	57.7 $\pm$ 5.4	88.6 $\pm$ 2.8	63.0 $\pm$ 9.2	30.5 $\pm$ 0.6	59.0 $\pm$ 0.5	63.8 $\pm$ 1.7	2.7
	GEBM	74.9 $\pm$ 8.6	62.1 $\pm$ 7.3	77.0 $\pm$ 4.0	55.7 $\pm$ 6.4	94.8 $\pm$ 1.9	55.5 $\pm$ 0.8	66.4 $\pm$ 1.8	2.6
Amazon Computers	EBM	53.7 $\pm$ 7.3	79.1 $\pm$ 4.5	56.6 $\pm$ 1.1	54.2 $\pm$ 3.0	43.5 $\pm$ 1.2	81.8 $\pm$ 1.8	60.0 $\pm$ 1.1	3.3
	Indep.	49.4 $\pm$ 4.0	63.9 $\pm$ 3.6	85.9 $\pm$ 2.7	73.1 $\pm$ 9.0	97.8 $\pm$ 4.1	49.5 $\pm$ 1.5	57.8 $\pm$ 0.7	3.9
	Local	48.3 $\pm$ 4.2	75.3 $\pm$ 7.2	50.7 $\pm$ 1.6	50.1 $\pm$ 2.0	57.0 $\pm$ 2.3	52.2 $\pm$ 0.5	67.3 $\pm$ 1.1	4.1
	Group	70.1 $\pm$ 3.7	77.4 $\pm$ 5.0	53.2 $\pm$ 1.6	51.8 $\pm$ 2.1	58.8 $\pm$ 2.0	49.2 $\pm$ 1.1	51.8 $\pm$ 1.8	4.3
	GEBM- E_θ	68.4 $\pm$ 4.8	81.3 $\pm$ 3.6	84.9 $\pm$ 1.9	75.9 $\pm$ 8.1	30.6 $\pm$ 0.9	49.9 $\pm$ 1.1	60.3 $\pm$ 0.7	2.7
	GEBM	65.0 $\pm$ 5.3	77.6 $\pm$ 6.2	79.5 $\pm$ 3.1	67.4 $\pm$ 7.1	96.0 $\pm$ 4.6	50.1 $\pm$ 1.2	61.3 $\pm$ 1.2	2.7
Coauthor-CS	EBM	89.6 $\pm$ 1.8	78.7 $\pm$ 3.3	69.4 $\pm$ 1.0	71.9 $\pm$ 2.4	37.2 $\pm$ 0.5	77.8 $\pm$ 1.2	60.9 $\pm$ 0.8	3.6
	Indep.	90.2 $\pm$ 0.8	65.7 $\pm$ 6.0	98.7 $\pm$ 0.3	100.0 $\pm$ 0.0	95.9 $\pm$ 3.1	61.5 $\pm$ 1.0	55.0 $\pm$ 0.5	3.0
	Local	62.0 $\pm$ 2.0	50.1 $\pm$ 3.9	54.1 $\pm$ 1.2	57.4 $\pm$ 1.7	61.7 $\pm$ 2.3	47.5 $\pm$ 0.1	77.1 $\pm$ 1.0	4.9
	Group	96.3 $\pm$ 0.9	80.8 $\pm$ 4.1	60.8 $\pm$ 1.8	64.2 $\pm$ 1.6	62.6 $\pm$ 1.7	52.8 $\pm$ 0.5	51.6 $\pm$ 1.2	3.9
	GEBM- E_θ	95.1 $\pm$ 1.1	88.6 $\pm$ 2.1	92.6 $\pm$ 0.9	96.0 $\pm$ 1.4	30.6 $\pm$ 0.4	58.1 $\pm$ 0.7	59.7 $\pm$ 0.5	3.0
	GEBM	93.4 $\pm$ 1.2	81.6 $\pm$ 5.1	93.9 $\pm$ 1.4	99.6 $\pm$ 0.2	92.8 $\pm$ 3.9	55.2 $\pm$ 0.8	65.7 $\pm$ 1.0	2.5
Coauthor-Physics	EBM	79.8 $\pm$ 3.2	92.0 $\pm$ 1.4	58.8 $\pm$ 1.2	62.1 $\pm$ 1.9	43.6 $\pm$ 0.9	77.3 $\pm$ 2.0	56.9 $\pm$ 0.8	3.4
	Indep.	63.9 $\pm$ 6.9	82.5 $\pm$ 2.8	90.6 $\pm$ 1.8	99.9 $\pm$ 0.1	83.9 $\pm$ 4.7	60.5 $\pm$ 1.2	56.3 $\pm$ 0.5	3.5
	Local	64.5 $\pm$ 4.8	84.2 $\pm$ 1.8	53.0 $\pm$ 0.7	54.5 $\pm$ 0.8	57.9 $\pm$ 1.3	50.9 $\pm$ 0.4	68.9 $\pm$ 1.1	4.5
	Group	88.2 $\pm$ 2.8	95.2 $\pm$ 0.7	54.0 $\pm$ 0.6	56.7 $\pm$ 0.7	56.2 $\pm$ 1.6	53.5 $\pm$ 0.9	53.6 $\pm$ 0.6	4.2
	GEBM- E_θ	89.2 $\pm$ 2.4	95.3 $\pm$ 0.9	74.8 $\pm$ 2.8	85.4 $\pm$ 2.9	31.0 $\pm$ 0.5	58.1 $\pm$ 1.3	56.8 $\pm$ 0.8	3.0
	GEBM	84.7 $\pm$ 3.8	95.4 $\pm$ 0.8	77.1 $\pm$ 2.7	94.1 $\pm$ 1.9	80.1 $\pm$ 4.4	55.9 $\pm$ 1.0	61.3 $\pm$ 0.8	2.3

Table 27: O.o.d-detection AUC-PR ($\uparrow$) using different EBMs in an inductive setting.

	Model	LoC (<i>last</i>)	LoC (<i>hetero.</i>)	Ber($\hat{p}$) (<i>near</i>)	Ber(0.5)	$\mathcal{N}(0, 1)$ (<i>far</i>)	Page Rank	Homophily	Rank($\downarrow$)
CoraML	EBM	87.5 $\pm$ 1.8	85.3 $\pm$ 2.0	68.7 $\pm$ 1.5	61.1 $\pm$ 4.9	24.7 $\pm$ 1.6	73.0 $\pm$ 1.7	67.9 $\pm$ 2.0	4.1
	Indep.	79.2 $\pm$ 1.4	76.9 $\pm$ 1.2	86.6 $\pm$ 2.1	98.5 $\pm$ 1.6	99.7 $\pm$ 0.3	55.9 $\pm$ 1.1	63.9 $\pm$ 1.5	3.7
	Local	86.9 $\pm$ 2.1	85.4 $\pm$ 1.5	56.5 $\pm$ 2.0	54.5 $\pm$ 2.2	62.1 $\pm$ 1.5	49.9 $\pm$ 0.7	79.8 $\pm$ 1.0	4.3
	Group	91.7 $\pm$ 1.2	89.6 $\pm$ 1.0	57.9 $\pm$ 1.0	61.3 $\pm$ 1.4	62.8 $\pm$ 2.1	48.1 $\pm$ 1.3	68.9 $\pm$ 1.5	3.3
	GEBM- E_θ	90.1 $\pm$ 1.6	88.0 $\pm$ 1.2	79.6 $\pm$ 2.3	76.9 $\pm$ 8.5	1.4 $\pm$ 0.3	52.3 $\pm$ 1.1	71.1 $\pm$ 2.3	3.1
	GEBM	89.7 $\pm$ 0.9	87.9 $\pm$ 1.4	78.9 $\pm$ 2.4	92.4 $\pm$ 3.7	89.8 $\pm$ 1.7	52.7 $\pm$ 1.0	73.1 $\pm$ 1.5	2.6
CoraML-LLM	EBM	88.2 $\pm$ 2.4	87.0 $\pm$ 2.3	n.a.	69.4 $\pm$ 2.4	34.1 $\pm$ 3.4	71.9 $\pm$ 2.1	69.0 $\pm$ 1.5	3.5
	Indep.	89.7 $\pm$ 1.5	79.7 $\pm$ 2.7	n.a.	91.9 $\pm$ 2.6	95.3 $\pm$ 2.4	51.2 $\pm$ 0.5	62.7 $\pm$ 2.1	3.4
	Local	89.7 $\pm$ 1.5	85.9 $\pm$ 2.6	n.a.	54.2 $\pm$ 1.9	63.0 $\pm$ 1.9	51.2 $\pm$ 0.5	79.7 $\pm$ 1.0	3.2
	Group	89.7 $\pm$ 1.5	91.0 $\pm$ 1.9	n.a.	57.5 $\pm$ 2.0	62.9 $\pm$ 2.1	51.2 $\pm$ 0.5	67.7 $\pm$ 2.0	3.2
	GEBM- E_θ	91.2 $\pm$ 1.5	90.6 $\pm$ 1.8	n.a.	88.9 $\pm$ 2.5	7.6 $\pm$ 0.9	51.0 $\pm$ 0.7	71.9 $\pm$ 2.2	2.9
	GEBM	89.7 $\pm$ 1.5	89.1 $\pm$ 2.6	n.a.	81.1 $\pm$ 3.6	86.0 $\pm$ 2.5	51.2 $\pm$ 0.5	72.9 $\pm$ 1.6	2.6
Citeseer	EBM	85.7 $\pm$ 1.2	85.5 $\pm$ 2.1	65.8 $\pm$ 2.0	55.0 $\pm$ 6.5	23.3 $\pm$ 1.2	66.3 $\pm$ 1.3	50.6 $\pm$ 1.0	4.0
	Indep.	76.3 $\pm$ 1.7	74.3 $\pm$ 2.4	69.2 $\pm$ 0.8	100.0 $\pm$ 0.0	93.5 $\pm$ 6.2	50.6 $\pm$ 1.4	49.2 $\pm$ 1.1	3.7
	Local	85.7 $\pm$ 2.2	89.1 $\pm$ 1.9	61.0 $\pm$ 1.8	72.7 $\pm$ 1.4	66.3 $\pm$ 2.1	47.4 $\pm$ 0.8	54.7 $\pm$ 0.5	3.2
	Group	88.7 $\pm$ 2.1	87.2 $\pm$ 2.4	55.7 $\pm$ 1.2	73.0 $\pm$ 1.3	66.2 $\pm$ 1.8	45.8 $\pm$ 0.9	51.6 $\pm$ 1.4	3.4
	GEBM- E_θ	88.4 $\pm$ 1.5	85.4 $\pm$ 2.2	64.7 $\pm$ 1.2	59.8 $\pm$ 9.0	5.5 $\pm$ 0.7	47.8 $\pm$ 0.9	51.3 $\pm$ 1.1	4.2
	GEBM	87.1 $\pm$ 2.2	86.7 $\pm$ 2.2	66.6 $\pm$ 1.2	98.5 $\pm$ 0.8	81.1 $\pm$ 2.8	48.0 $\pm$ 1.1	51.9 $\pm$ 1.1	2.5
PubMed	EBM	66.1 $\pm$ 3.0	62.8 $\pm$ 4.5	n.a.	58.5 $\pm$ 7.5	26.4 $\pm$ 0.9	71.8 $\pm$ 2.7	53.4 $\pm$ 0.9	3.9
	Indep.	64.0 $\pm$ 1.8	59.6 $\pm$ 2.9	n.a.	89.9 $\pm$ 5.5	75.2 $\pm$ 11.7	50.8 $\pm$ 0.8	53.5 $\pm$ 0.8	3.4
	Local	66.7 $\pm$ 3.3	65.2 $\pm$ 3.7	n.a.	57.5 $\pm$ 2.3	60.9 $\pm$ 1.4	47.8 $\pm$ 0.7	61.6 $\pm$ 2.0	3.5
	Group	69.3 $\pm$ 3.8	70.4 $\pm$ 3.8	n.a.	59.6 $\pm$ 2.1	58.9 $\pm$ 2.0	51.0 $\pm$ 0.9	53.4 $\pm$ 2.4	2.5
	GEBM- E_θ	68.4 $\pm$ 3.6	65.4 $\pm$ 5.1	n.a.	66.7 $\pm$ 10.3	8.0 $\pm$ 0.4	49.4 $\pm$ 1.0	59.9 $\pm$ 1.0	3.1
	GEBM	68.4 $\pm$ 3.0	66.3 $\pm$ 3.7	n.a.	84.0 $\pm$ 5.1	73.0 $\pm$ 6.2	49.6 $\pm$ 0.8	57.1 $\pm$ 1.9	2.4
Amazon Photo	EBM	75.2 $\pm$ 4.7	75.9 $\pm$ 3.7	58.5 $\pm$ 1.5	51.5 $\pm$ 1.4	41.3 $\pm$ 1.6	90.5 $\pm$ 0.9	56.8 $\pm$ 1.4	5.1
	Indep.	75.5 $\pm$ 3.8	76.2 $\pm$ 3.0	86.8 $\pm$ 3.1	59.7 $\pm$ 8.9	93.8 $\pm$ 3.8	58.2 $\pm$ 1.6	62.2 $\pm$ 1.9	3.1
	Local	76.9 $\pm$ 7.3	91.4 $\pm$ 2.2	71.6 $\pm$ 16.9	56.3 $\pm$ 7.4	77.5 $\pm$ 17.5	54.5 $\pm$ 0.6	62.2 $\pm$ 1.9	3.6
	Group	89.7 $\pm$ 3.4	81.5 $\pm$ 5.2	72.7 $\pm$ 15.5	57.0 $\pm$ 6.8	77.5 $\pm$ 17.6	53.4 $\pm$ 1.1	62.2 $\pm$ 1.9	3.5
	GEBM- E_θ	85.5 $\pm$ 5.9	82.1 $\pm$ 3.2	93.6 $\pm$ 0.6	66.3 $\pm$ 9.1	0.5 $\pm$ 0.1	60.7 $\pm$ 1.0	61.6 $\pm$ 1.8	2.9
	GEBM	84.2 $\pm$ 5.3	87.4 $\pm$ 3.9	83.3 $\pm$ 3.9	58.4 $\pm$ 8.0	92.0 $\pm$ 2.7	57.8 $\pm$ 1.2	62.2 $\pm$ 1.9	2.8
Amazon Computers	EBM	75.3 $\pm$ 4.3	84.5 $\pm$ 3.2	57.8 $\pm$ 1.6	54.6 $\pm$ 2.3	40.8 $\pm$ 1.4	87.2 $\pm$ 1.5	60.6 $\pm$ 0.8	4.1
	Indep.	77.3 $\pm$ 1.5	79.5 $\pm$ 1.9	89.5 $\pm$ 1.8	80.7 $\pm$ 4.5	94.2 $\pm$ 8.2	50.9 $\pm$ 1.5	55.9 $\pm$ 1.0	4.1
	Local	77.1 $\pm$ 4.3	89.5 $\pm$ 1.9	52.2 $\pm$ 0.7	50.5 $\pm$ 0.7	57.9 $\pm$ 1.5	53.7 $\pm$ 0.6	68.2 $\pm$ 1.3	3.6
	Group	89.7 $\pm$ 4.2	91.1 $\pm$ 1.2	53.8 $\pm$ 1.5	52.9 $\pm$ 1.5	58.8 $\pm$ 1.8	51.3 $\pm$ 1.0	52.4 $\pm$ 2.5	3.6
	GEBM- E_θ	85.1 $\pm$ 3.3	89.1 $\pm$ 2.7	91.7 $\pm$ 1.5	81.8 $\pm$ 5.0	0.2 $\pm$ 0.1	52.7 $\pm$ 1.1	63.7 $\pm$ 0.6	2.7
	GEBM	84.8 $\pm$ 2.8	90.1 $\pm$ 1.5	84.3 $\pm$ 2.8	72.9 $\pm$ 5.3	92.0 $\pm$ 7.7	52.1 $\pm$ 1.1	61.9 $\pm$ 1.5	2.9
Coauthor-CS	EBM	88.6 $\pm$ 1.1	91.8 $\pm$ 1.9	68.7 $\pm$ 0.9	70.9 $\pm$ 1.3	29.4 $\pm$ 0.8	81.5 $\pm$ 1.2	59.5 $\pm$ 0.3	3.8
	Indep.	89.5 $\pm$ 0.9	88.9 $\pm$ 3.2	99.1 $\pm$ 0.2	100.0 $\pm$ 0.0	97.6 $\pm$ 1.9	63.2 $\pm$ 0.9	55.6 $\pm$ 0.7	3.0
	Local	73.1 $\pm$ 1.8	85.1 $\pm$ 2.9	57.6 $\pm$ 0.5	58.1 $\pm$ 0.9	63.9 $\pm$ 1.6	45.6 $\pm$ 0.4	77.8 $\pm$ 0.7	4.9
	Group	96.5 $\pm$ 1.1	93.8 $\pm$ 2.7	61.7 $\pm$ 1.6	64.0 $\pm$ 1.6	64.5 $\pm$ 1.4	51.8 $\pm$ 0.9	54.5 $\pm$ 1.9	3.8
	GEBM- E_θ	94.2 $\pm$ 1.1	96.4 $\pm$ 1.0	95.3 $\pm$ 0.6	97.5 $\pm$ 0.5	0.4 $\pm$ 0.1	59.1 $\pm$ 0.9	63.7 $\pm$ 0.7	2.8
	GEBM	92.7 $\pm$ 1.3	93.6 $\pm$ 2.9	96.0 $\pm$ 0.6	99.2 $\pm$ 0.5	95.1 $\pm$ 2.0	56.0 $\pm$ 0.7	68.4 $\pm$ 1.1	2.7
Coauthor-Physics	EBM	91.5 $\pm$ 1.7	95.3 $\pm$ 1.6	59.7 $\pm$ 1.2	62.0 $\pm$ 1.4	39.6 $\pm$ 0.8	83.0 $\pm$ 1.0	56.5 $\pm$ 0.7	3.8
	Indep.	88.2 $\pm$ 1.8	92.9 $\pm$ 1.1	92.9 $\pm$ 1.3	99.5 $\pm$ 0.8	81.5 $\pm$ 4.8	66.4 $\pm$ 0.4	54.9 $\pm$ 0.3	3.7
	Local	89.7 $\pm$ 1.7	93.8 $\pm$ 2.2	53.9 $\pm$ 0.6	55.7 $\pm$ 1.1	58.8 $\pm$ 0.8	52.4 $\pm$ 0.3	67.9 $\pm$ 1.5	4.5
	Group	94.8 $\pm$ 2.1	97.8 $\pm$ 0.7	56.1 $\pm$ 0.9	57.8 $\pm$ 1.1	58.3 $\pm$ 1.0	55.2 $\pm$ 0.4	56.6 $\pm$ 1.3	3.6
	GEBM- E_θ	95.7 $\pm$ 1.0	97.6 $\pm$ 1.1	82.5 $\pm$ 2.7	90.0 $\pm$ 2.8	4.0 $\pm$ 0.6	63.0 $\pm$ 0.6	58.7 $\pm$ 0.9	2.8
	GEBM	93.7 $\pm$ 1.9	97.2 $\pm$ 1.2	83.8 $\pm$ 2.7	93.4 $\pm$ 2.7	78.6 $\pm$ 3.7	60.9 $\pm$ 0.6	61.9 $\pm$ 1.0	2.7

Table 28: O.o.d-detection AUC-ROC ($\uparrow$) using different EBMs in a transductive setting.

Model		LoC (<i>last</i>)	LoC (<i>hetero.</i>)	Ber($\hat{p}$) (<i>near</i>)	Ber(0.5)	$\mathcal{N}(0, 1)$ (<i>far</i>)	Page Rank	Homophily	Rank($\downarrow$)
CoraML	EBM	83.0 $\pm$ 2.1	69.6 $\pm$ 4.3	66.1 $\pm$ 3.1	56.3 $\pm$ 6.2	35.0 $\pm$ 1.4	70.2 $\pm$ 2.9	62.5 $\pm$ 2.1	3.6
	Indep.	72.2 $\pm$ 1.0	55.5 $\pm$ 2.3	82.1 $\pm$ 4.3	98.1 $\pm$ 2.6	99.8 $\pm$ 0.2	56.3 $\pm$ 1.4	56.6 $\pm$ 1.4	3.7
	Local	78.0 $\pm$ 4.5	64.7 $\pm$ 3.5	53.8 $\pm$ 2.2	52.4 $\pm$ 2.7	61.9 $\pm$ 3.3	54.1 $\pm$ 0.4	76.0 $\pm$ 1.0	4.0
	Group	88.0 $\pm$ 1.9	72.0 $\pm$ 2.7	55.5 $\pm$ 3.3	60.0 $\pm$ 3.9	61.7 $\pm$ 3.7	50.5 $\pm$ 0.6	57.6 $\pm$ 1.4	3.8
	GEBM- E_θ	86.5 $\pm$ 2.2	71.6 $\pm$ 3.1	71.7 $\pm$ 3.8	67.5 $\pm$ 10.3	30.1 $\pm$ 1.5	52.8 $\pm$ 0.9	61.4 $\pm$ 2.2	3.6
	GEBM	86.1 $\pm$ 1.6	73.2 $\pm$ 2.8	71.2 $\pm$ 4.4	91.6 $\pm$ 5.2	91.7 $\pm$ 1.8	53.3 $\pm$ 0.7	64.3 $\pm$ 1.4	2.4
CoraML-LLM	EBM	84.0 $\pm$ 4.7	73.0 $\pm$ 4.5	n.a.	66.0 $\pm$ 4.4	41.0 $\pm$ 2.8	69.6 $\pm$ 2.5	64.7 $\pm$ 1.6	3.3
	Indep.	85.9 $\pm$ 2.9	58.8 $\pm$ 5.3	n.a.	82.6 $\pm$ 4.1	97.2 $\pm$ 1.4	52.5 $\pm$ 0.4	55.3 $\pm$ 2.5	3.7
	Local	85.9 $\pm$ 2.9	64.0 $\pm$ 4.6	n.a.	52.5 $\pm$ 3.7	61.9 $\pm$ 3.0	52.5 $\pm$ 0.4	77.2 $\pm$ 0.9	3.3
	Group	85.9 $\pm$ 2.9	75.0 $\pm$ 5.3	n.a.	55.9 $\pm$ 3.3	61.7 $\pm$ 3.3	52.5 $\pm$ 0.4	57.7 $\pm$ 1.6	3.5
	GEBM- E_θ	88.3 $\pm$ 3.2	78.3 $\pm$ 4.9	n.a.	84.8 $\pm$ 4.2	31.7 $\pm$ 1.5	52.6 $\pm$ 0.7	62.6 $\pm$ 1.9	2.1
	GEBM	85.9 $\pm$ 2.9	74.6 $\pm$ 6.1	n.a.	72.4 $\pm$ 5.3	88.4 $\pm$ 2.5	52.5 $\pm$ 0.4	65.2 $\pm$ 1.9	2.7
Citeseer	EBM	66.3 $\pm$ 4.4	72.1 $\pm$ 4.2	60.8 $\pm$ 4.4	50.0 $\pm$ 5.0	35.3 $\pm$ 2.5	60.2 $\pm$ 1.8	51.3 $\pm$ 0.7	3.4
	Indep.	48.6 $\pm$ 2.2	57.7 $\pm$ 3.5	62.7 $\pm$ 2.2	100.0 $\pm$ 0.0	96.2 $\pm$ 3.6	50.0 $\pm$ 1.5	49.6 $\pm$ 0.8	3.7
	Local	60.4 $\pm$ 4.6	78.7 $\pm$ 3.8	58.6 $\pm$ 3.4	73.4 $\pm$ 3.5	65.8 $\pm$ 2.7	48.3 $\pm$ 0.9	54.7 $\pm$ 1.1	2.9
	Group	67.0 $\pm$ 5.4	71.9 $\pm$ 3.5	52.8 $\pm$ 2.6	74.4 $\pm$ 2.5	64.5 $\pm$ 2.4	45.7 $\pm$ 0.8	51.4 $\pm$ 0.5	3.9
	GEBM- E_θ	68.8 $\pm$ 4.6	69.8 $\pm$ 3.2	56.4 $\pm$ 2.7	51.7 $\pm$ 5.9	31.2 $\pm$ 2.0	47.2 $\pm$ 0.6	51.3 $\pm$ 0.4	4.4
	GEBM	65.8 $\pm$ 5.4	74.1 $\pm$ 3.9	59.1 $\pm$ 2.5	98.6 $\pm$ 0.8	83.3 $\pm$ 2.5	47.6 $\pm$ 1.0	51.4 $\pm$ 0.3	2.9
PubMed	EBM	53.9 $\pm$ 2.9	28.6 $\pm$ 4.0	n.a.	53.1 $\pm$ 5.2	38.0 $\pm$ 0.4	64.4 $\pm$ 4.1	55.1 $\pm$ 1.3	3.2
	Indep.	50.5 $\pm$ 1.8	24.3 $\pm$ 1.6	n.a.	90.3 $\pm$ 5.6	84.9 $\pm$ 7.5	50.1 $\pm$ 0.7	52.5 $\pm$ 0.8	3.4
	Local	50.6 $\pm$ 2.3	27.7 $\pm$ 2.8	n.a.	56.4 $\pm$ 2.6	60.2 $\pm$ 2.0	47.2 $\pm$ 0.4	59.6 $\pm$ 1.6	3.7
	Group	58.8 $\pm$ 3.6	38.2 $\pm$ 5.4	n.a.	58.9 $\pm$ 2.7	57.4 $\pm$ 1.8	49.7 $\pm$ 0.8	50.1 $\pm$ 1.7	2.6
	GEBM- E_θ	56.5 $\pm$ 3.7	34.3 $\pm$ 6.3	n.a.	57.7 $\pm$ 7.5	31.3 $\pm$ 0.4	48.4 $\pm$ 0.6	57.5 $\pm$ 0.7	2.9
	GEBM	55.9 $\pm$ 3.2	30.8 $\pm$ 3.5	n.a.	83.3 $\pm$ 6.2	77.3 $\pm$ 5.3	48.4 $\pm$ 0.6	54.3 $\pm$ 1.5	2.9
Amazon Photo	EBM	64.3 $\pm$ 6.3	49.1 $\pm$ 4.3	59.3 $\pm$ 2.5	52.0 $\pm$ 2.1	44.3 $\pm$ 1.2	89.5 $\pm$ 1.4	58.9 $\pm$ 1.9	3.8
	Indep.	66.4 $\pm$ 5.6	48.3 $\pm$ 4.3	81.1 $\pm$ 5.3	52.8 $\pm$ 6.4	95.8 $\pm$ 2.6	55.8 $\pm$ 1.6	58.6 $\pm$ 2.0	3.4
	Local	62.5 $\pm$ 9.9	75.5 $\pm$ 5.9	67.1 $\pm$ 13.6	52.1 $\pm$ 3.3	79.4 $\pm$ 18.3	55.3 $\pm$ 0.7	58.6 $\pm$ 2.0	3.9
	Group	80.0 $\pm$ 6.2	50.0 $\pm$ 8.5	68.1 $\pm$ 12.4	52.1 $\pm$ 3.6	79.4 $\pm$ 18.3	51.3 $\pm$ 0.6	58.6 $\pm$ 2.0	3.7
	GEBM- E_θ	78.3 $\pm$ 7.1	52.3 $\pm$ 5.4	90.8 $\pm$ 1.6	59.9 $\pm$ 8.9	30.7 $\pm$ 0.7	57.6 $\pm$ 0.8	58.6 $\pm$ 2.4	2.7
	GEBM	77.1 $\pm$ 8.1	63.2 $\pm$ 10.1	75.7 $\pm$ 4.9	51.9 $\pm$ 5.3	94.4 $\pm$ 1.7	55.3 $\pm$ 0.9	58.6 $\pm$ 2.0	3.5
Amazon Computers	EBM	50.9 $\pm$ 6.4	77.3 $\pm$ 4.4	57.4 $\pm$ 2.1	53.3 $\pm$ 2.9	43.7 $\pm$ 1.2	84.4 $\pm$ 2.5	59.4 $\pm$ 0.9	3.8
	Indep.	49.6 $\pm$ 2.6	63.9 $\pm$ 2.9	84.1 $\pm$ 2.0	70.5 $\pm$ 4.9	96.7 $\pm$ 4.8	49.2 $\pm$ 1.5	55.3 $\pm$ 0.9	4.1
	Local	49.2 $\pm$ 4.3	79.8 $\pm$ 3.3	50.5 $\pm$ 1.3	49.0 $\pm$ 1.2	57.3 $\pm$ 1.9	51.8 $\pm$ 0.5	67.7 $\pm$ 1.4	3.9
	Group	72.8 $\pm$ 3.2	79.9 $\pm$ 2.2	53.0 $\pm$ 1.7	51.7 $\pm$ 1.6	58.5 $\pm$ 2.2	48.9 $\pm$ 0.7	49.3 $\pm$ 2.2	4.0
	GEBM- E_θ	62.6 $\pm$ 6.1	79.9 $\pm$ 3.6	87.9 $\pm$ 1.9	73.6 $\pm$ 5.9	30.3 $\pm$ 0.5	50.0 $\pm$ 0.9	59.6 $\pm$ 0.7	2.6
	GEBM	66.8 $\pm$ 3.5	80.4 $\pm$ 3.4	77.4 $\pm$ 2.9	64.4 $\pm$ 4.9	94.7 $\pm$ 4.9	49.7 $\pm$ 1.1	59.5 $\pm$ 1.5	2.6
Coauthor-CS	EBM	87.8 $\pm$ 1.7	72.3 $\pm$ 4.4	69.5 $\pm$ 1.3	72.2 $\pm$ 1.9	38.0 $\pm$ 0.6	77.5 $\pm$ 1.6	60.5 $\pm$ 0.4	3.6
	Indep.	88.7 $\pm$ 1.2	58.9 $\pm$ 7.6	98.8 $\pm$ 0.3	99.9 $\pm$ 0.1	98.7 $\pm$ 1.0	60.7 $\pm$ 0.8	53.0 $\pm$ 0.6	3.0
	Local	61.0 $\pm$ 1.7	46.9 $\pm$ 4.1	55.2 $\pm$ 0.5	56.3 $\pm$ 0.8	63.3 $\pm$ 1.8	47.6 $\pm$ 0.2	77.8 $\pm$ 0.8	4.9
	Group	96.5 $\pm$ 0.8	74.8 $\pm$ 8.2	60.8 $\pm$ 1.6	63.4 $\pm$ 1.6	63.6 $\pm$ 1.6	52.9 $\pm$ 0.7	48.8 $\pm$ 1.4	3.9
	GEBM- E_θ	94.7 $\pm$ 0.8	84.7 $\pm$ 3.7	92.8 $\pm$ 1.0	96.3 $\pm$ 0.8	31.2 $\pm$ 0.5	57.5 $\pm$ 0.7	59.3 $\pm$ 0.7	3.0
	GEBM	92.8 $\pm$ 0.9	74.9 $\pm$ 7.5	94.9 $\pm$ 0.7	99.2 $\pm$ 0.6	96.4 $\pm$ 1.4	54.9 $\pm$ 0.6	64.7 $\pm$ 1.3	2.5
Coauthor-Physics	EBM	75.3 $\pm$ 3.5	88.0 $\pm$ 4.0	59.4 $\pm$ 1.7	61.1 $\pm$ 1.9	43.3 $\pm$ 0.6	78.4 $\pm$ 1.4	57.5 $\pm$ 0.6	3.6
	Indep.	60.5 $\pm$ 4.6	79.4 $\pm$ 2.7	91.5 $\pm$ 1.6	99.1 $\pm$ 1.7	88.9 $\pm$ 2.8	61.1 $\pm$ 0.6	55.5 $\pm$ 0.4	3.2
	Local	57.3 $\pm$ 4.8	74.5 $\pm$ 7.7	53.0 $\pm$ 0.7	53.9 $\pm$ 0.9	58.5 $\pm$ 1.0	51.1 $\pm$ 0.1	70.5 $\pm$ 1.0	4.8
	Group	86.4 $\pm$ 3.0	93.7 $\pm$ 2.0	54.7 $\pm$ 1.4	56.8 $\pm$ 1.5	57.3 $\pm$ 0.9	54.2 $\pm$ 0.5	53.5 $\pm$ 1.0	3.9
	GEBM- E_θ	86.6 $\pm$ 1.8	93.1 $\pm$ 3.2	76.3 $\pm$ 3.5	84.4 $\pm$ 4.7	30.7 $\pm$ 0.3	58.8 $\pm$ 0.8	57.6 $\pm$ 0.6	2.8
	GEBM	80.6 $\pm$ 3.7	91.5 $\pm$ 4.3	78.6 $\pm$ 3.5	90.9 $\pm$ 4.1	84.4 $\pm$ 2.7	56.7 $\pm$ 0.4	61.7 $\pm$ 0.8	2.7

Table 29: O.o.d-detection AUC-PR ($\uparrow$) using different EBMs in a transductive setting.

	Model	LoC	Ber($\hat{p}$)	$\mathcal{N}(0, 1)$	Homo.
APNP	EBM	90.5	71.9	17.0	73.3
	Safe	91.6	55.9	33.1	73.6
	GEBM	88.6	67.9	67.9	78.8
GCN $\sigma = 0.5$	EBM	60.3	52.2	17.9	52.2
	Safe	58.7	50.0	31.9	50.8
	GEBM	14.1	35.5	99.6	32.8
GCN $\sigma = 1.0$	EBM	75.4	60.6	3.6	61.6
	Safe	74.4	53.9	20.0	61.6
	GEBM	29.9	37.5	99.6	40.4
GCN $\sigma = 2.0$	EBM	84.1	75.4	2.6	67.2
	Safe	89.8	60.1	17.4	69.1
	GEBM	89.4	61.1	99.6	75.9

Table 30: O.o.d. detection AUC for APNP and GCN with spectral normalization at different weight scales σ .

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: We identify three main contributions in the introduction Section 1. Our novel model is described in Section 4.3, the formal proofs regarding integrability of EBMs and Gaussian regularization are given in Sections 4.2 and 4.3 and a novel evidential interpretation of our EBM framework can be found in Section 4.4. We support the claimed efficacy of our model at different structural scales by its strong out-of-distribution detection performance on a comprehensive suite of distribution shifts (Appendix C).

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: We discuss limitations in Section 6: We only aim at epistemic uncertainty estimation and can not improve on aleatoric metrics. Our method is tailored toward node classification in a homophilic setting, which is also acknowledged. Our method is less effective on centrality-based splits, which is discussed in Sections 5.2 and 6 and in-depth in Appendix C.6.

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: We explicitly state the assumptions that we deal with piecewise affine classifiers and provide formal proofs for all statements in Appendix A. For the core theoretical contribution, we provide intuition in Section 4.2.

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: We provide code to reproduce our experiments online. Furthermore, we additionally also list all hyperparameters for all baselines and our model explicitly in Appendix B.

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: We provide code to reproduce our experiments.

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: Experimental details, entailing dataset splits, architecture choices, optimization, etc., are provided in Appendix B. Additionally, they are listed as default configurations in the public code.

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: We omit error-bars and standard deviations in the main text for clarity, but supply them in Appendix C. They are explicitly referred to as standard deviations.

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: Details about compute resources and a discussion about memory and training time are provided in Appendix B.

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: Our research aims to improve reliability of GNNs. All datasets are commonly used benchmarks that are publicly available. Our paper fully conforms with the code of ethics.

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: We aim to contribute to making AI more reliable by improving on uncertainty quantification. We discuss the impact of our work beyond commonly known risks and concerns in research on AI in Section 6.

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: We do not believe that model has an increased risk for misuse. We encourage users for active evaluation of the uncertainty provided in Section 6.

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 cite papers producing the commonly used node classification benchmark. We downloaded all datasets from PyTorch Geometric, as described in Appendix B. We report licenses where we could find them. All datasets have been used in open research for several years and form a well-established benchmark.

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: [\[NA\]](#)

Justification: Our paper does not provide new assets.

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: Our paper does not involve crowdsourcing nor 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: Our paper does not involve crowdsourcing nor research with human subjects.

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.