

CREDAL GRAPH NEURAL NETWORKS FOR ROBUST UNCERTAINTY QUANTIFICATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Uncertainty quantification is essential for deploying reliable Graph Neural Networks (GNNs), where existing approaches primarily rely on Bayesian inference or ensembles. In this paper, we introduce the first credal graph neural networks (CGNNs), which extend credal learning to the graph domain by training GNNs to output set-valued predictions in the form of credal sets. To account for the distinctive nature of message passing in GNNs, we develop a complementary approach to credal learning that leverages different aspects of layer-wise information propagation. We assess our approach on uncertainty quantification in node classification under out-of-distribution conditions. Our analysis highlights the critical role of the graph homophily assumption in shaping the effectiveness of uncertainty estimates. Extensive experiments demonstrate that CGNNs deliver more reliable representations of epistemic uncertainty and achieve state-of-the-art performance under distributional shift on heterophilic graphs.

1 INTRODUCTION

In safety-critical applications, a machine learning (ML) model must not only be accurate but also aware of its own limitations. The central challenge in making ML models reliable for real-world deployment is particularly acute for Graph Neural Networks (GNNs), where the inherent dependencies between data points violate the standard i.i.d. assumption. In a classical ML model, its confidence is often conflated into a single predictive probability, which fails to distinguish different sources of uncertainty (Gawlikowski et al., 2023). This lack of specificity is critical on graphs, where a model can be uncertain about a node’s class either because its local neighborhood is genuinely ambiguous (a property of the data) or because the node and its connections are entirely novel and outside the training distribution (a property of the model’s knowledge) (Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021). To build more robust systems, we must formally disentangle the two primary sources of uncertainty: *aleatoric uncertainty*, the inherent and often irreducible randomness in the data-generating process, and *epistemic uncertainty*, which stems from the model’s limited knowledge and is potentially reducible with additional data or improved models (Gal & Ghahramani, 2016; Valdenegro-Toro & Mori, 2022). This decomposition underpins many modern approaches for uncertainty-aware learning and OOD detection (Kendall & Gal, 2017).

The Bayesian framework provides a natural and powerful lens for uncertainty disentanglement (Malinin & Gales, 2018). The predictive posterior distribution, which gives the probability vector $\mathbf{y}$ for a new input $\mathbf{x}$ given the training data $\mathcal{D}$, is formulated by marginalizing over the entire space of plausible model parameters $\boldsymbol{\theta}$:

$$P(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \int \underbrace{P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}_{\text{Aleatoric}} \cdot \underbrace{p(\boldsymbol{\theta}|\mathcal{D})}_{\text{Epistemic}} d\boldsymbol{\theta}. \quad (1)$$

The integral in Equation 1 explicitly separates the two uncertainty sources. The likelihood term, $P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, represents the aleatoric uncertainty. It captures the stochasticity of the output for a single, fixed model configuration $\boldsymbol{\theta}$. Even if we knew the true model parameters perfectly, this term would remain, as it models the inherent ambiguity in the data. The posterior distribution over the model’s parameters, $p(\boldsymbol{\theta}|\mathcal{D})$, represents the epistemic uncertainty. It quantifies the model’s lack of knowledge by describing a set of plausible parameter configurations consistent with the observed

data $\mathcal{D}$. When data is scarce, this distribution is wide and diffuse, reflecting high epistemic uncertainty; as more data is observed we expect it to sharpen and concentrates around the optimal parameters, thereby reducing the model’s ignorance. The final predictive distribution is thus an average of many possible models, weighted by their posterior probability, which correctly accounts for both what the model has learned and what it still does not know.

The most principled approach to uncertainty quantification, Bayesian Neural Networks (BNNs) (Goan & Fookes, 2020), places probability measures over network parameters, thereby defining a probabilistic model over models. This results in predictive outputs described by second-order probabilities, which capture uncertainty over all possible likelihoods. However, such methods incur substantial computational overhead during both training and inference (Wang et al., 2024a); further discussion is provided in Appendix D.

Despite the progress of existing approaches (see Section 2), a critical gap remains. Most current frameworks, particularly those based on message passing or uncertainty diffusion, implicitly assume graph homophily, i.e., that connected nodes tend to share similar features or labels (Ma et al.). As a result, their performance and the reliability of their uncertainty estimates deteriorate significantly in heterophilic settings, a challenge only recently beginning to be systematically addressed (Fuchslugruber et al., 2025). This highlights the need for new, robust uncertainty quantification frameworks that do not rely on homophily and can operate effectively across diverse graph structures.

Main Contributions. We address the problem of uncertainty estimation in graph learning task, by proposing an approach based on credal learning, and assessing specifically the underlooked aspect of graph heterophily. Our main contributions are: (i) we introduce *Credal Graph Neural Networks (CGNNs)*, the first framework that extends credal learning to the graph domain; (ii) we develop a credal graph learning architecture that leverages layer-wise message passing to provide more faithful uncertainty representations; and (iii) we conduct extensive experiments on both homophilic and heterophilic benchmarks, showing that CGNNs achieve state-of-the-art performance in out-of-distribution detection and highlight the critical role of graph structure in shaping uncertainty estimates.

2 RELATED WORKS

Quantifying uncertainty in Graph Neural Networks (GNNs) (Bacciu et al., 2020) is a relevant and rapidly developing research direction, with several methods being proposed recently to capture the reliability of GNN predictions (Wang et al., 2024a; Chen et al., 2024). The inherent dependencies among nodes in a graph introduce unique challenges, leading to a diverse landscape of quantification strategies. Existing approaches can be broadly organized into three main families, each with distinct computational and modeling characteristics.

The most computationally efficient family consists of *Single Deterministic Models*. The simplest methods in this category are post-hoc heuristics applied to a standard GNN’s output, such as using the maximum softmax probability as a measure of confidence or calculating the predictive entropy. While easy to implement, these approaches provide a single, undifferentiated measure of uncertainty and are often sensitive to model miscalibration (Guo et al., 2017). A more principled deterministic approach is evidential deep learning, where the GNN learns to output the parameters of a higher-order Dirichlet distribution. This allows for directly modeling uncertainty over the categorical output space in a single forward pass, providing a richer characterization of the predictive uncertainty (Zhao et al.; Stadler et al.).

A second major paradigm is *Bayesian Graph Neural Networks (Bayesian GNNs)*, which adapt the Bayesian framework specifically to graph-structured data. In these models, one places priors not only over standard network parameters (e.g., weights) but also over graph-specific components such as edge connectivity or node feature propagation. For example, adaptive connection sampling (Hasanzadeh et al., 2020) treats edges (or adjacency masks) probabilistically, and inference techniques like Monte Carlo Dropout (Gal & Ghahramani, 2016) or Variational Inference (Hoffman et al., 2013) are used to sample over both weights and possible graph instantiations during prediction. This both reflects uncertainty in what the graph structure (and features) imply, and in the model’s parameters. Works such as Munikoti et al. (2023) illustrate how Bayesian methods in graphs model uncertainty over graph structure, node features, and edge sampling, not just traditional weights. This

provides a theoretically grounded framework for capturing uncertainty stemming from the model’s own parameters and from uncertainty in the graph data (structure or features), but at the cost of higher computational complexity (Jia et al., 2020) and the challenge of defining suitable priors for graph structures.

Bridging the gap between theoretical rigor and practical scalability, *Ensemble Methods* have become a powerful and popular alternative. This approach averages the predictions from multiple GNNs that are trained independently. Uncertainty is then quantified by measuring the disagreement or variance among the predictions of the individual models in the ensemble (Mallick et al., 2022; Busk et al., 2023). While empirically powerful and often outperforming more complex Bayesian methods, ensembles also carry a significant computational and memory burden, as they require training and storing multiple full models.

3 CREDAL GRAPH NEURAL NETWORKS

This section introduces our Credal Graph Neural Network (CGNN) framework. We begin by outlining the principles of credal learning, which models uncertainty using sets of probability distributions to disentangle its aleatoric and epistemic sources (Section 3.1). We then present our novel architecture for Credal Graph Learning in Section 3.2. Our primary model, `CredalLJ`, leverages a joint latent representation from all GNN layers, and we introduce two variants for comparison: `Credal Final` and the post-hoc `Credal Ensemble`. Finally, Section 3.3 describes the training procedure, which is based on a Distributionally Robust Optimization (DRO) objective to ensure robustness to distributional shifts.

3.1 CREDAL LEARNING

We begin by providing the basics of definitions behind our credal-based approach to uncertainty in graph learning. Credal learning has been proposed to address the limitations of BNNs and Ensemble-based approaches by explicitly modeling uncertainty as a convex set of probability distributions, known as a credal set $\mathcal{P}$ (Caprio et al., 2024). In practice, the model constructs this set by outputting a probability interval for each class, with each interval being defined by a lower and an upper bound. Figure 1 provides a visual example for 3-class classification task. The shaded area represents the credal set, which can be interpreted as the region of the hypothesis that is coherent with the training data.

Instead of averaging distributions, as needed by Bayesian Model Averaging or Ensembles (details in Appendix E), uncertainty can be defined by the size and location of $\mathcal{P}$ inside the simplex. Similarly to Wang et al. (2024b), we quantify this using generalized entropy, which defines the total uncertainty TU and the aleatoric uncertainty AU as the maximum and minimum possible Shannon entropy for any distribution p within the credal set (Abellán et al., 2006):

$$\text{TU} := \overline{H}(\mathcal{P}) = \max_{p \in \mathcal{P}} H(p), \quad \text{AU} := \underline{H}(\mathcal{P}) = \min_{p \in \mathcal{P}} H(p). \quad (2)$$

The epistemic uncertainty is the difference, $\text{EU} := \overline{H}(\mathcal{P}) - \underline{H}(\mathcal{P})$, which intuitively depends on the volume of the credal set and on its position inside the simplex. A confident prediction (Figure 1(a)) corresponds to a small set near a vertex. High data uncertainty (Figure 1(b)) is a small set near the center. High epistemic uncertainty is a large set covering a significant portion of the simplex (Figure 1 (c and d)), indicating that many different probability distributions are considered plausible.

3.2 CREDAL GRAPH LEARNING

Credal graph layer. To move beyond single-point probability predictions, we build on the Credal Layer proposed by Wang et al. (2024c) as a replacement for the classification layer of a GNN. Let $\mathbf{z}_v$ be a generic embedding of node v (as computed by some layer of a GNN), the Credal Layer would then take $\mathbf{z}_v$ as input and, for each of the C classes, output two values: an interval midpoint m_v^c and a half-length $h_v^c \geq 0$. Collecting these for all classes gives two vectors $\mathbf{m}_v = (m_v^1, \dots, m_v^C)$ and $\mathbf{h}_v = (h_v^1, \dots, h_v^C)$, each of length C . This is achieved using the transformation:

$$\mathbf{m}_v = g(W \times \mathbf{z}_v + \mathbf{b}), \quad \mathbf{h}_v = g'(W' \times \mathbf{z}_v + \mathbf{b}'), \quad (3)$$

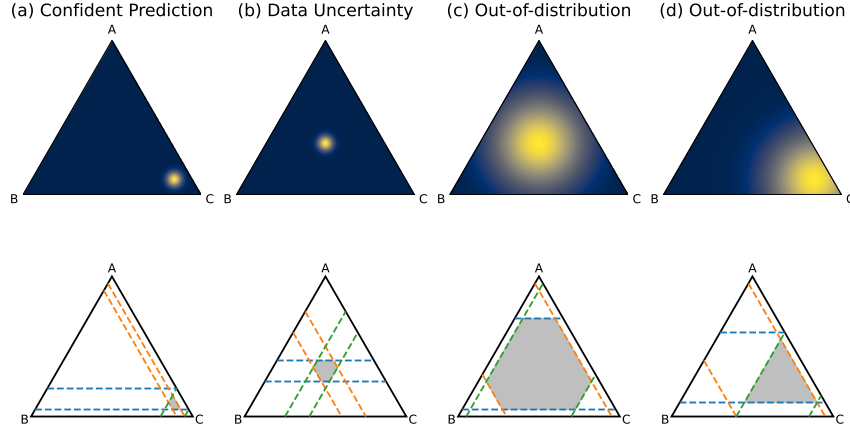


Figure 1: Visualization of aleatoric uncertainty (AU) and epistemic uncertainty (EU) for a 3-class classification problem, corresponding to the disentanglement in Equation 1. The top row shows a Bayesian continuous representation, while the bottom row shows the corresponding credal set representation. (a) Low AU and low EU: the model is confident in its prediction. (b) High AU, low EU: the model attributes uncertainty to noisy data. (c) High AU and high EU: the model faces out-of-distribution data. (d) Low AU, high EU: the model encounters novel data and attributes uncertainty to its parameters rather than to noise. In the credal set representation (bottom row), the shaded regions inside the simplex correspond to lower and upper probability bounds for each class.

where g and g' are activation functions, with $g' \geq 0$. These values define an interval for each class, $[a_v^L, a_v^U] := [m_v - h_v, m_v + h_v]$. To transform these into valid lower and upper probability bounds, $[q_v^L, q_v^U]$, that satisfy the necessary convexity conditions, a specialized Interval SoftMax activation (Wang et al., 2025) is applied:

$$q_v^{L_i} = \frac{\exp(a_v^{L_i})}{\exp(a_v^{L_i}) + \sum_{k \neq i} \exp(a_v^{U_k})}, \quad q_v^{U_i} = \frac{\exp(a_v^{U_i})}{\exp(a_v^{U_i}) + \sum_{k \neq i} \exp(a_v^{L_k})}. \quad (4)$$

The resulting probability intervals define a credal set $\mathcal{P}_v = \{\mathbf{q} | q^i \in [q_v^{L_i}, q_v^{U_i}], \sum_{i=1}^C q^i = 1\}$, which constitutes the final credal prediction for node v .

Information propagation on graphs. A fundamental question is which representation $\mathbf{z}_v$ to use for a node v . For models processing independent and identically distributed (i.i.d.) data, the layer-wise processing can be described by the Data Processing Inequality (DPI) (Polyanskiy & Wu, 2017). This principle states that the mutual information between a layer’s latent representation Z^l and the target label Y cannot increase with network depth: $I(Y; Z^{l+1}) \leq I(Y; Z^l)$. Information is progressively filtered and compressed, and any information about Y lost at layer i cannot be recovered later.

The mechanism of information propagation within GNNs differs fundamentally from that in standard feed-forward networks. The DPI model of information flow is insufficient for GNNs, where the message-passing mechanism violates the i.i.d. assumption by design. A node’s v representation at layer l , namely $\mathbf{z}_v^l$, is updated using information from its neighbors, effectively expanding its receptive field with each layer according to:

$$\mathbf{z}_v^l = \Phi^l(\mathbf{z}_v^{l-1}, \Psi(\{\Omega^l(\mathbf{z}_q^{l-1}) \mid q \in N_v\})). \quad (5)$$

In this formula, N_v is the set of neighbors of node v , their representation in the layer $l-1$ is projected into a new space via the function Ω and aggregated into a single neighborhood representation using a permutation-invariant function Ψ (such as max, mean, or sum). Finally, the function Φ combines this aggregated neighborhood representation with the embedding of node v to produce the updated representation.

Fuchsgruber et al. (2025) formalize this with a *Data Processing Equality for Message Passing Neural Networks*, which shows that the information about a node’s label Y in its representation Z^{l+1} at layer $l + 1$ decomposes as:

$$I(Y; Z^{l+1}) = I(Y; Z^l) - \Delta_-^{[0:l]} + \Delta_+^{l+1}. \quad (6)$$

Here $I(Y; Z^l)$ is the information about the target Y contained in the node’s own representation at the previous layer l . $\Delta_-^{[0:l]}$ is the *relative information loss or recovery term*. It quantifies how much information about Y from the node’s ego-graph up to l hops is lost (while also accounting for the recovery of previously lost information retained in node’s neighbors) during the $l + 1$ -th message-passing step. Δ_+^{l+1} is the *information gain term*, representing new, additional information about Y that is incorporated from the newly included $l + 1$ -hop neighbors.

This decomposition is key to understanding the difficulty of graph GNNs in particular in heterophilic learning, as observed by many authors (Micheli & Tortorella, 2023; Liang et al., 2023; Ma et al., 2021; Zhu et al., 2020; Luan et al., 2024; Lim et al., 2021). In homophilic settings, neighbors are semantically similar, so the information gain Δ_+^{l+1} is often low (i.e. information tends to be redundant). In contrast, in heterophilic settings, neighbors are semantically *different*, meaning the information gain Δ_+^{l+1} can be substantial at each layer. Each latent representation z_v^l thus provides unique and crucial information. The fundamental flaw of standard message-passing in this regime is that it acts as a local smoothing operator, indiscriminately mixing the central node’s signal with these diverse neighboring signals. This corruption leads to representations of different classes becoming indistinguishable.

Motivated by these observations we propose to use in input to the credal layer z_v^{joint} , that is the joint latent representation of a node v , computed by concatenating the node’s embeddings from all layers:

$$z_v^{\text{joint}} = [z_v^0 || z_v^1 || \dots || z_v^L], \quad (7)$$

where z_v^0 are the initial (input) node features.

By using z_v^{joint} as node’s v embedding, we ensure the credal prediction is conditioned on all the information the GNN has processed, allowing for a more faithful and robust estimation of both the prediction and its associated uncertainty. The full architecture, named `CredalLJ` for Latent Joint, is depicted in Figure 2.

To isolate the effect of the credal output layer in a more standard GNN setting, we evaluate a variant where it is applied only to the final hidden layer. In this model, which we term `Credal Final`, we remove the concatenation of embeddings from all layers. The credal output layer, takes only the final latent representation, z_v^L , of the GNN backbone as its input.

Post-hoc credal graph learning. The credal graph learning approach outlined above requires to modify the original GNN architecture, hence it cannot be applied post-hoc to already trained model. As an alternative method, we also explore a post-hoc approach for generating credal predictions from a set of GNNs trained for a standard classification task. Inspired by (Wang et al., 2024b), we define a `Credal Ensemble` leveraging an ensemble of vanilla GNNs that requires no changes to their training. For a given input node v , each model h_m produces a standard probabilistic prediction in the form of a softmax vector. The core principle of the `Credal Ensemble` is to interpret the set of these predictions, $\{\mathbf{p}_1, \dots, \mathbf{p}_M\}$, as the vertices of a credal set. The final credal prediction for node v , $\mathcal{P}_v$, is therefore defined as the convex hull of the ensemble’s outputs:

$$\mathcal{P}_v = \text{Conv}(\{\mathbf{p}_1, \dots, \mathbf{p}_M\}). \quad (8)$$

This resulting set contains all possible convex combinations of the individual model predictions, thereby capturing the epistemic uncertainty expressed through the disagreement among the ensemble members. This method offers a straightforward way to obtain a credal prediction from any existing pre-trained GNN ensemble.

3.3 TRAINING

To train a model capable of output predictions in the form of a credal set we frame our problem within the setting of learning from a collection of data sets $\{\mathcal{D}_1, \dots, \mathcal{D}_Q\}$, $\mathcal{D}_i =$

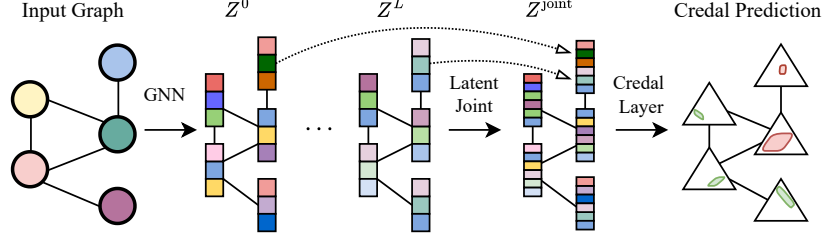


Figure 2: Overview of the proposed credal prediction framework. An input graph is processed by a Graph Neural Network (GNN), which iteratively aggregates node representations across layers ($Z^0, \dots, Z^L$), gaining and losing information at each step. The latent joint representation Z^{joint} allows recovery of the full trajectory of each node across layers. A credal layer then maps Z^{joint} to a credal prediction, providing an informative representation of uncertainty. Green credal sets correspond to reliable predictions, while red credal sets indicate nodes that are likely out-of-distribution (OOD) due to high aleatoric or epistemic uncertainty.

$\{(x_{i,1}, y_{i,1}), \dots, (x_{i,n_i}, y_{i,n_i})\}$, where each $\mathcal{D}_i$ is issued from a different “domain” characterized by its own unknown data-generating probability distribution (Caprio et al., 2024). The idea is that by leveraging the available evidence $\mathcal{D}_i$ the model is able to elicit a credal set that contains the true data generating process for a new set of data $\mathcal{D}_{Q+1}$. In practice, however, we are often limited to a single training set $\mathcal{D}$, which can be viewed as an aggregation of data from these various underlying domains. We adopt, following Wang et al. (2024c), a strategy inspired by Distributionally Robust Optimization (DRO) (Ben-Tal et al., 2013). Instead of minimizing only the classical empirical risk, the objective is to also find model parameters θ that minimize the worst-case risk over an uncertainty set of distributions $\mathcal{U}$ around the empirical distribution $\hat{P}$:

$$\min_{\theta \in \Theta} \left\{ \sup_{U \in \mathcal{U}} \mathbb{E}_{(x,y) \sim U} [\mathcal{L}(h_{\theta}, (x, y))] \right\}, \quad (9)$$

where $\mathcal{L}$ is a given loss function and h_{θ} is the hypothesis parameterized by θ . Training the model to be robust against this internally-generated adverse distribution compels it to produce predictions in the form of a credal set, represented by a collection of probability intervals.

Following heuristics (Huang et al., 2022; Oren et al., 2019), the DRO loss can be approximated as:

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \text{CE}(q_n^U, y_n) + \frac{1}{\delta N} \sum_{n \in \mathcal{H}_{\delta}} \text{CE}(q_n^L, y_n), \quad (10)$$

where

$$\mathcal{H}_{\delta} = \arg \top_{\delta N} \left\{ \text{CE}(q_n^L, y_n) \mid n = 1, \dots, N \right\}, \quad (11)$$

N is the number of training points, CE is the cross-entropy and δ is an hyper-parameter that indicates the proportion of difficult training examples to consider for an estimate of test distribution divergence. The first component of Equation 10 relies on the training distribution as is, and therefore tends to encourage more *optimistic* or upper-bound predictions for the class scores (q^U). In contrast, the second component emphasizes the boundary cases and training outliers by operating on q^L , simulating potential shifts between training and test distributions. This encourages the model to produce more *pessimistic* or lower-bound predictions. As a consequence, the width between q^U and q^L reflects the model’s ignorance about how much the future test distribution may differ from the training distribution. By leveraging the outlier cases observed at training time we approximate the DRO loss (Equation 9), allowing the model to better estimate the uncertainty it may encounter at test time.

4 EXPERIMENTAL ANALYSIS

In this section, we empirically validate our proposed Credal Graph Neural Networks. We first introduce the datasets used for our evaluation (Section 4.1) and define the experimental setting for

out-of-distribution (OOD) node detection in Section 4.2. We then describe the baselines and ablation studies used for comparison (Sections 4.3 and 4.4), and finally, present and analyze the results in Section 4.5.

4.1 DATASETS

We evaluate our method on a diverse suite of six benchmark graph datasets, encompassing various domains and structural properties, as summarized in Table 1.

Table 1: Statistics and OOD splits for the benchmark datasets. Validation and test set sizes are reported as in-distribution (ID) / out-of-distribution (OOD).

Dataset	# Nodes	# Edges	# Classes	Homophily	OOD Classes	ID Classes	Train (ID)	Val (ID/OOD)	Test (ID/OOD)
Chameleon	2,277	31,421	5	He	{0,1}	{2,3,4}	647	438 / 291	276 / 180
Squirrel	5,201	198,493	5	He	{0,1}	{2,3,4}	1,515	982 / 682	622 / 419
ArXiv	169,343	1,166,243	5	He	{0,1}	{2,3,4}	69,523	23,415 / 10,453	23,245 / 10,625
Patents	2,923,922	13,975,788	5	He	{0,1}	{2,3,4}	1,053,055	351,259 / 233,525	350,730 / 234,055
Coauthor	18,333	163,788	15	Ho	{0,...,3}	{4,...,14}	8,813	2,907 / 759	2,964 / 704
Reddit2	232,965	23,213,838	41	Ho	{0,...,10}	{11,...,40}	109,517	17,343 / 6,356	40,601 / 14,733

The *Coauthors* dataset is a computer science co-authorship network where nodes represent authors, connected by an edge if they have co-authored a paper (Shchur et al., 2018). Node features are derived from paper keywords, and the task is to predict each author’s primary field of study. The *Chameleon* and *Squirrel* datasets are Wikipedia networks where nodes are web pages linked by hyperlinks (Rozemberczki et al., 2021). Their features are derived from informative nouns on each page, and the classification task is to categorize pages based on their average monthly traffic. *Reddit2* is a dataset constructed from Reddit posts, where nodes represent individual posts with features generated from text embeddings (Hamilton et al., 2017). An edge connects two posts if the same user has commented on both, and the prediction task is to identify the subreddit to which each post belongs. *ArXiv* is a citation network where nodes are academic papers and directed edges represent citations (Hu et al., 2020; Ma et al.). Node features are embeddings of the paper’s title and abstract, and the objective is to predict the publication year. Finally, *Patents* is a citation network of U.S. utility patents, where each node is a patent and edges indicate citations between them (Leskovec et al., 2005; Lim et al., 2021; Ma et al.). Features are derived from patent metadata, and the task is to predict the patent’s grant date.

4.2 SETTING

We address the problem of node-level out-of-distribution (OOD) detection in a transductive setting using a *Leave-Out-Class* strategy during training. Let $G = (\mathcal{V}, \mathcal{E}, X)$ denote a single attributed graph, where $\mathcal{V}$ is the set of nodes, $\mathcal{E}$ is the set of edges, and $X \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the node feature matrix. The set of all node classes is partitioned into in-distribution (ID) classes $\mathcal{C}_{ID}$ and out-of-distribution (OOD) classes $\mathcal{C}_{OOD}$, such that $\mathcal{C}_{ID} \cap \mathcal{C}_{OOD} = \emptyset$.

The node set $\mathcal{V}$ is divided into training, validation, and test subsets. During training, nodes belonging to OOD classes in $\mathcal{C}_{OOD}$ are masked, i.e., their labels are treated as unknown. Consequently, the model is trained to solve a classification problem over only the ID classes $\mathcal{C}_{ID}$. Importantly, in this setting, the true labels of test nodes may include both ID and OOD classes, i.e., the test classes form a superset of the training classes.

The objective is to learn a model that classify the in-distribution nodes into their respective classes within $\mathcal{C}_{ID}$ and produce a high uncertainty score for nodes belonging to the OOD classes.

4.3 BASELINES

We evaluate our proposed credal learning approaches against a suite of established baselines for uncertainty estimation and OOD detection (Ma et al.). These include a family of widely-used post-hoc methods that operate on a single pre-trained GNN, such as the *Energy*-based score, which is calculated from the pre-softmax logits (Liu et al., 2020); *ODIN*, which applies temperature scaling and input perturbations (Liang et al., 2017); and the *Mahalanobis* baseline, which measures the distance of a test sample’s latent representation from the training data’s class-conditional distributions (Lee et al., 2018). Similarly, we employ a *K-Nearest Neighbors* (KNN) approach,

where uncertainty is derived from the average latent-space distance to the nearest training samples (Sun et al., 2022). We also compare against a method specifically designed for graph data, namely GNNSafe, a graph-specific Energy-Based Model that incorporates a label propagation mechanism (Wu et al., 2023).

However, the de-facto reference model against which new uncertainty quantification methods are measured is the `Classical Ensemble`. Despite its conceptual simplicity, this approach consistently achieves state-of-the-art performance across a wide range of tasks and datasets, making it the target method to outperform (Lakshminarayanan et al., 2017; Gustafsson et al., 2020; Ovadia et al., 2019; Abe et al., 2022). Its strength, lies in its combination of high performance with practical simplicity: it is easy to implement, scales effectively, and is largely hyperparameter-free, requiring only the independent training of multiple standard models.

Additional details on model selection are reported in Appendix C.

4.4 ABLATED MODELS

To better understand the effect of each component in our proposed `CredalLJ` GNN model, we conduct a series of ablation studies.

The `Credal Final` describes in Section 3.2 removes the effect of the joint trajectory of the embedding so that the credal module is constrained to rely solely on the final node representation.

To test whether the performance of our model stems primarily from its latent joint representation rather than the credal module itself, we conduct a second ablation. We replace our credal output layer with a simple KNN density estimator, a method inspired by the Joint Latent Density Estimation (JLDE) principle recently proposed by Fuchsgruber et al. (2025). Specifically, we compute the OOD score based on the k-Nearest Neighbors (KNN) distance within the joint latent space, z_v^{joint} , used by our full model. We refer to this model as `KNNLJ`.

Finally, the `Credal Ensemble` introduced in Section 3.2 removes the specialized training procedure for the credal module, aiming to estimate the credal set from a set of standard prediction.

4.5 RESULTS

The OOD detection performance of our proposed methods and all baselines is presented in Table 2, with results reported in AUROC. The experiments span six diverse datasets, including both homophilic and highly heterophilic graphs, allowing for a comprehensive evaluation of robustness to different graph structures. Additional results are reported in Appendix B.

Table 2: OOD detection performance across all datasets (AUROC $\uparrow$). For models that allow uncertainty disentanglement, both aleatoric and epistemic uncertainty scores are reported (aleatoric/epistemic). Best results are highlighted in **bold**, second-best results are underlined, and third-best results are shown with a light gray background. Results with variance are reported in Appendix B.

Method	Chameleon	Squirrel	ArXiv	Patents	Coauthor	Reddit2
Energy	58.25	44.47	50.16	43.33	94.47	43.93
KNN	56.81	52.79	56.86	52.55	88.23	65.63
ODIN	57.98	47.28	48.16	43.21	94.17	43.09
Mahalanobis	51.82	53.79	59.52	58.72	82.49	68.98
GNNSafe	50.42	35.88	35.30	27.35	94.82	61.99
Classical ensemble	74.00/30.22	58.32/59.13	58.20/ 65.45	48.23/60.35	95.30 /94.81	58.84/ 72.09
KNNLJ	70.06	71.06	45.35	46.74	32.37	36.55
Credal final	76.29/67.27	75.02/65.85	64.65/65.66	68.92/69.73	74.80/50.74	70.57/69.35
Credal ensemble	74.98 /29.54	59.03/53.66	58.05/50.97	47.41/ 64.64	93.85/93.72	61.16/57.72
CredalLJ	72.37/ 77.67	73.89/ 77.04	65.77 /63.79	70.78 /60.06	86.58/70.88	66.35/67.31

The results highlight the limitations of existing uncertainty quantification methods, particularly on heterophilic graphs. Our `CredalLJ` achieves the best performance on all heterophilic benchmarks

(Chameleon, Squirrel, ArXiv, Patents), while remaining competitive on homophilic ones, being only a few percentage points below the `Classical Ensemble`, which stands out as the best baseline.

The `Credal Final` consistently ranks second on all heterophilic benchmarks, being only 1–2 percentage points below the `CredalLJ`, which highlights the importance of the `Credal` module.

`Credal Ensemble` and `KNNLJ` performed worse than expected. In particular, `KNNLJ` proved to be an unreliable uncertainty estimator, achieving good performance only on two datasets, while `Credal Ensemble` slightly underperformed the `Classical Ensemble`, with the exception of the Patents dataset. This suggests that a simple post-hoc application of credal theory offers limited gains without architectural and training modifications.

As expected, the `Classical Ensemble` outperformed all baselines on both homophilic and heterophilic datasets. However, the baselines maintained competitiveness in the homophilic setting.

On the homophilic Coauthors dataset, many standard post-hoc methods, including `Energy`, and the graph-specific `GNNSafe`, achieve strong performance. Similarly for `Reddit2` where Mahalanobis and KNN are competitive. However, the performance of nearly all single-model baselines collapses dramatically on the heterophilic datasets, especially for `GNNSafe` which relies on homophilic smoothing kernels.

A deeper analysis of the results reveals several intriguing patterns. First, the optimal type of uncertainty for OOD detection is highly dataset-dependent. For our state-of-the-art `CredalLJ` model, epistemic uncertainty provides the strongest signal on the smaller heterophilic graphs, whereas aleatoric uncertainty is more effective on the larger ones. This suggests that as the scale of the graph increases, the most reliable OOD indicator may shift from the model’s own ignorance (epistemic) to the inherent ambiguity of the data itself (aleatoric).

We observe that, in general, models rely on different uncertainty types for different graphs. This highlights the practical value of our proposed methods, which can successfully disentangle and leverage both aleatoric and epistemic uncertainty to adapt to a wide variety of graph properties and complexities.

This interplay between aleatoric and epistemic uncertainty in the context of OOD detection has been noted in prior work (Stadler et al., 2021; Fuchsluger et al., 2024; 2025) and requires further study.

5 CONCLUSIONS

In this paper, we introduced `Credal Graph Neural Networks`, the first framework to extend credal learning to the graph domain for robust uncertainty quantification. Our approach addresses the limitations of existing methods, which often falter on graphs that violate the homophily assumption. We proposed a novel architecture that leverages a joint latent representation of the full information propagation trajectory across all GNN layers. This, combined with a training objective inspired by Distributionally Robust Optimization, allows our models to produce set-valued predictions in the form of credal sets, enabling a principled disentanglement of aleatoric and epistemic uncertainty.

Our experiments demonstrate the effectiveness of our approach, particularly on challenging heterophilic graphs. The proposed framework sets a new state-of-the-art for out-of-distribution detection across all heterophilic benchmarks, where the performance of most single-model baselines collapses dramatically. Furthermore, our models remain highly competitive on homophilic datasets, showcasing their robustness and versatility across diverse graph structures. These results highlight the importance of both the novel joint latent architecture and the specialized credal training procedure for achieving reliable uncertainty estimates.

By introducing the first credal GNNs, this work opens a new and promising research direction for uncertainty-aware graph machine learning. Our findings also underscore the need for a deeper investigation into the complex interplay between aleatoric and epistemic uncertainty in the context of OOD detection. Future work should therefore focus on better understanding this relationship, exploring how these distinct sources of uncertainty can be optimally combined or selectively leveraged to build more reliable and trustworthy GNNs for safety-critical applications.

Reproducibility statement To ensure the reproducibility of our research, we have made our complete source code and experimental setup publicly available. The repository, which includes the implementation of our proposed Credal GNN and all other benchmarked models, can be accessed at the following anonymous link: <https://anonymous.4open.science/r/CGNN-ICLR26/>. The repository contains detailed instructions for setting up the required environment and scripts to replicate the experiments presented in this paper. Furthermore, all datasets utilized in our study are established, open-source benchmarks from the graph machine learning community. These datasets can be retrieved from their original publications, which are cited in our manuscript where each dataset is first introduced. We believe that the provided code and the public nature of the datasets offer sufficient resources for the research community to verify our findings and build upon our work.

Ethics Statement The research conducted in this paper conforms in every aspect with the ICLR Code of Ethics. Our study does not involve human subjects, sensitive personal data, or applications with foreseeable harmful consequences. No ethical concerns are anticipated regarding data usage, methodology, or findings.

REFERENCES

- Taiga Abe, Estefany Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John P Cunningham. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660, 2022.
- Joaquín Abellán, George J Klir, and Serafín Moral. Disaggregated total uncertainty measure for credal sets. *International Journal of General Systems*, 35(1):29–44, 2006.
- Davide Bacciu, Federico Errica, Alessio Micheli, and Marco Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.
- Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- Jonas Busk, Mikkel N Schmidt, Ole Winther, Tejs Vegge, and Peter Bjørn Jørgensen. Graph neural network interatomic potential ensembles with calibrated aleatoric and epistemic uncertainty on energy and forces. *Physical Chemistry Chemical Physics*, 25(37):25828–25837, 2023.
- Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. *Advances in Neural Information Processing Systems*, 37:38665–38694, 2024.
- Chao Chen, Chenghua Guo, Rui Xu, Xiangwen Liao, Xi Zhang, Sihong Xie, Hui Xiong, and Philip Yu. Uncertainty quantification on graph learning: A survey. *arXiv preprint arXiv:2404.14642*, 2024.
- Dominik Fuchsguber, Tom Wollschläger, and Stephan Günnemann. Energy-based epistemic uncertainty for graph neural networks. *Advances in Neural Information Processing Systems*, 37:34378–34428, 2024.
- Dominik Fuchsguber, Tom Wollschläger, Johannes Bordne, and Stephan Günnemann. Uncertainty estimation for heterophilic graphs through the lens of information theory. *arXiv preprint arXiv:2505.22152*, 2025.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.
- Jakob Gawlikowski, Cedrique Rovile Njiteucheu 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.
- Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pp. 45–87. Springer, 2020.

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 318–319, 2020.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.
- Arman Hasanzadeh, Ehsan Hajiramezanali, Shahin Boluki, Mingyuan Zhou, Nick Duffield, Krishna Narayanan, and Xiaoning Qian. Bayesian graph neural networks with adaptive connection sampling. In *International conference on machine learning*, pp. 4094–4104. PMLR, 2020.
- Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *the Journal of machine Learning research*, 14(1):1303–1347, 2013.
- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.
- Zeyi Huang, Haohan Wang, Dong Huang, Yong Jae Lee, and Eric P Xing. The two dimensions of worst-case training and their integrated effect for out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9631–9641, 2022.
- 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.
- Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pp. 548–557. PMLR, 2022.
- Xiaotao Jia, Jianlei Yang, Runze Liu, Xueyan Wang, Sorin Dan Cotofana, and Weisheng Zhao. Efficient computation reduction in bayesian neural networks through feature decomposition and memorization. *IEEE transactions on neural networks and learning systems*, 32(4):1703–1712, 2020.
- Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.
- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 177–187, 2005.
- Langzhang Liang, Xiangjing Hu, Zenglin Xu, Zixing Song, and Irwin King. Predicting global label relationship matrix for graph neural networks under heterophily. *Advances in Neural Information Processing Systems*, 36:10909–10921, 2023.
- Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.
- Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in neural information processing systems*, 34:20887–20902, 2021.

- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.
- Sitao Luan, Chenqing Hua, Qincheng Lu, Liheng Ma, Lirong Wu, Xinyu Wang, Minkai Xu, Xiao-Wen Chang, Doina Precup, Rex Ying, et al. The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges. *arXiv preprint arXiv:2407.09618*, 2024.
- Longfei Ma, Yiyu Sun, Kaize Ding, Zemin Liu, and Fei Wu. Revisiting score propagation in graph out-of-distribution detection. 37:4341–4373. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/08342dc6ab69f23167b4123086ad4d38-Abstract-Conference.html.
- Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. Is homophily a necessity for graph neural networks? *arXiv preprint arXiv:2106.06134*, 2021.
- Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.
- Tanwi Mallick, Prasanna Balaprakash, and Jane Macfarlane. Deep-ensemble-based uncertainty quantification in spatiotemporal graph neural networks for traffic forecasting. *arXiv preprint arXiv:2204.01618*, 2022.
- Alessio Micheli and Domenico Tortorella. Addressing heterophily in node classification with graph echo state networks. *Neurocomputing*, 550:126506, 2023.
- Sai Munikoti, Deepesh Agarwal, Laya Das, and Balasubramaniam Natarajan. A general framework for quantifying aleatoric and epistemic uncertainty in graph neural networks. *Neurocomputing*, 521:1–10, 2023.
- Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pp. 211–249. Springer, 2017.
- Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.
- Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.
- Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18033–18048. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/95b431e51fc53692913da5263c214162-Abstract.html.
- 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.
- Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.
- Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE, 2022.

- 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*, 2024a.
- Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation on out-of-distribution detection. *arXiv preprint arXiv:2405.15047*, 2024b.
- Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, Hans Hallez, et al. Credal deep ensembles for uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:79540–79572, 2024c.
- Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, 185:107198, 2025.
- Larry Wasserman. Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107, 2000.
- 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.
- Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc. URL <https://proceedings.neurips.cc/paper/2020/hash/968c9b4f09cbb7d7925f38aea3484111-Abstract.html>.
- Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

A LLM USAGE

Large Language Models (LLMs) were used as general-purpose assistive tools during the preparation of this paper. Their role was limited to two aspects: (i) improving the writing quality (e.g., grammar correction, rephrasing for clarity, and suggesting minor improvements to text structure), and (ii) assisting with retrieval and discovery of related work (e.g., suggesting potentially relevant references for further manual inspection by the authors).

All content generated or suggested by LLMs was carefully reviewed, verified, and, where necessary, edited by the authors to ensure accuracy and faithfulness to the intended scientific contributions. LLMs were not used to generate scientific content, including research ideas, hypothesis formulation, experimental design, data analysis, or interpretation of results.

The authors take full responsibility for all parts of the paper.

B ADDITIONAL RESULTS

Table 3: In-distribution (ID) test classification performance (F1-score $\uparrow$). For our credal models, we report results based on two predictions, using the vectors q_L and q_U . We report mean $\pm$ standard deviation over 5 runs. Best results are highlighted in **bold**, second-best results are underlined, and third-best results are shown with a light gray background.

Method	Chameleon	Squirrel	ArXiv	Patents	Coauthor	Reddit2
Vanilla GNN	57.61 $\pm$ 0.98	46.46 $\pm$ 2.25	<u>51.81 $\pm$ 0.90</u>	53.35 $\pm$ 0.43	95.23 $\pm$ 1.41	79.47 $\pm$ 0.36
Credal Final (q_L)	53.99 $\pm$ 1.42	39.55 $\pm$ 0.74	51.03 $\pm$ 0.40	51.20 $\pm$ 0.41	50.53 $\pm$ 0.89	27.38 $\pm$ 0.36
Credal Final (q_U)	55.43 $\pm$ 0.38	39.55 $\pm$ 0.43	51.03 $\pm$ 0.42	51.94 $\pm$ 0.40	28.60 $\pm$ 0.81	24.63 $\pm$ 0.36
CredalLJ (q_L)	42.75 $\pm$ 0.41	39.71 $\pm$ 0.45	52.39 $\pm$ 0.39	49.34 $\pm$ 0.40	75.44 $\pm$ 0.98	23.83 $\pm$ 0.37
CredalLJ (q_U)	36.96 $\pm$ 0.39	36.17 $\pm$ 1.44	51.01 $\pm$ 0.40	49.34 $\pm$ 0.41	43.76 $\pm$ 0.42	27.29 $\pm$ 0.37

We report the in-distribution (ID) classification performance, measured by the F1-score, in Table 3. The results show that the Vanilla GNN baseline generally outperforms our credal models on ID data. This outcome is expected and highlights a known trade-off in robust machine learning: the Vanilla GNN is optimized solely for empirical risk minimization, maximizing performance on data from the training distribution. Our credal models, in contrast, are regularized via a Distributionally Robust Optimization (DRO) objective to enhance out-of-distribution (OOD) robustness, which naturally results in more conservative predictions and a slight decrease in ID performance.

Notably, the performance degradation on ID classification appears correlated with the number of classes. The degradation is most pronounced on datasets with a high number of classes, such as Reddit2 (41 classes) and Coauthor (15 classes), while the performance is much more competitive on the 5-class datasets like ArXiv. This suggests that deriving a single, accurate point prediction from a credal set becomes inherently more challenging as the dimensionality of the target space grows. While methods exist to address this, such as using an Intersection Probability to map a credal set to a single probability (Wang et al., 2024b), or employing Probability Interval Dimension Reduction (PIDR) to manage the computational complexity in many-class settings (Wang et al., 2024c), our focus in this work remains on the quality of the uncertainty representation itself rather than on optimizing the point-prediction decision rule.

The OOD detection results with standard deviation are presented in Table 4.

C MODEL SELECTION

To ensure a fair and robust comparison, we performed extensive hyperparameter tuning for all baseline and proposed models, with the search spaces detailed in Table 5. For the Vanilla GNN and our Credal models (Final Layer and Latent Joint), we employed a Bayesian optimization strategy with 30 trials for each dataset and model combination. The Vanilla GNN was optimized to maximize the validation F1-score, reflecting its primary goal of in-distribution accuracy. In contrast, our Credal models were optimized to maximize the validation AUROC based on their epistemic uncertainty

Table 4: OOD detection performance across all datasets (AUROC $\uparrow$). We report the mean and standard deviation over 5 runs. For models that allow uncertainty disentanglement, we report aleatoric (AU) and epistemic (EU) scores separately.

Method	Chameleon	Squirrel	ArXiv	Patents	Coauthor	Reddit2
Energy	58.25 $\pm$ 4.63	44.47 $\pm$ 1.62	50.16 $\pm$ 4.71	43.33 $\pm$ 3.48	94.47 $\pm$ 0.45	43.93 $\pm$ 1.24
KNN	56.81 $\pm$ 6.28	52.79 $\pm$ 1.55	56.86 $\pm$ 1.39	52.55 $\pm$ 1.24	88.23 $\pm$ 2.39	65.63 $\pm$ 1.50
ODIN	57.98 $\pm$ 2.46	47.28 $\pm$ 2.05	48.16 $\pm$ 2.37	43.21 $\pm$ 0.98	94.17 $\pm$ 0.22	43.09 $\pm$ 0.26
Mahalanobis	51.82 $\pm$ 3.31	53.79 $\pm$ 0.75	59.52 $\pm$ 1.47	58.72 $\pm$ 1.25	82.49 $\pm$ 0.47	68.98 $\pm$ 1.21
GNNSafe	50.42 $\pm$ 0.86	35.88 $\pm$ 0.67	35.30 $\pm$ 0.34	27.35 $\pm$ 0.27	94.82 $\pm$ 0.98	61.99 $\pm$ 0.96
Classical ensemble (AU)	74.00 $\pm$ 0.96	58.32 $\pm$ 1.19	58.20 $\pm$ 0.93	48.23 $\pm$ 0.93	95.30 $\pm$ 1.15	58.84 $\pm$ 1.46
Classical ensemble (EU)	30.22 $\pm$ 1.23	59.13 $\pm$ 0.86	65.45 $\pm$ 1.07	60.35 $\pm$ 0.86	94.81 $\pm$ 1.47	72.09 $\pm$ 0.86
KNNLJ	70.06 $\pm$ 0.50	71.06 $\pm$ 0.50	45.35 $\pm$ 0.70	46.74 $\pm$ 1.09	32.37 $\pm$ 1.07	36.55 $\pm$ 0.83
Credal ensemble (AU)	74.98 $\pm$ 1.23	59.03 $\pm$ 0.86	58.05 $\pm$ 0.50	47.41 $\pm$ 0.83	93.85 $\pm$ 1.46	61.16 $\pm$ 0.50
Credal ensemble (EU)	29.54 $\pm$ 0.70	53.66 $\pm$ 1.07	50.97 $\pm$ 0.70	64.64 $\pm$ 1.09	93.72 $\pm$ 0.70	57.72 $\pm$ 0.50
Credal final (AU)	79.29 $\pm$ 0.30	75.02 $\pm$ 6.20	64.65 $\pm$ 0.20	68.92 $\pm$ 2.74	74.80 $\pm$ 6.97	70.57 $\pm$ 1.47
Credal final (EU)	67.27 $\pm$ 2.32	65.85 $\pm$ 0.57	65.66 $\pm$ 1.27	66.73 $\pm$ 0.23	50.74 $\pm$ 1.36	69.35 $\pm$ 2.24
CredalLJ (AU)	72.37 $\pm$ 7.16	73.89 $\pm$ 0.47	65.77 $\pm$ 0.21	70.78 $\pm$ 7.34	86.58 $\pm$ 1.43	66.35 $\pm$ 0.11
CredalLJ (EU)	77.67 $\pm$ 1.32	77.04 $\pm$ 8.25	63.79 $\pm$ 0.08	60.06 $\pm$ 0.47	70.88 $\pm$ 1.92	67.31 $\pm$ 0.69

Table 5: Hyperparameter search spaces for the proposed models. $U(a, b)$ denotes a uniform distribution between a and b .

Hyperparameter	Vanilla GNN	Classical & Credal Ensemble	Credal Final & LJ	KNNLJ
Learning Rate (lr)	$U(10^{-5}, 10^{-1})$	—	$U(10^{-5}, 10^{-1})$	—
Hidden Channels	{64, 128, 256}	—	{64, 128, 256}	—
Num Layers	{2, 3}	—	{2, 3}	—
Weight Decay	$U(10^{-7}, 10^{-1})$	—	$U(10^{-7}, 10^{-1})$	—
GNN Type	{GCN, SAGE}	—	{GCN, SAGE}	—
Ensemble Size (M)	—	{2, ..., 15}	—	—
Delta (δ)	—	—	$U(0.5, 1.0)$	—
k	—	—	—	{5, 10, 20, 50, 100, 200}

score, directly tuning them for OOD detection performance. The Classical/Credal Ensemble baseline was constructed following a two-stage process to ensure its strength. First, we trained a large pool of 100 Vanilla GNNs using an extended Bayesian search. From this pool, we selected the top-performing models (based on their validation F1-score) to construct ensembles of varying sizes M . For the KNNLJ we used the best Vanilla model to compute the embeddings.

For the baseline models, we used the hyperparameter range in their original papers.

D BAYESIAN MODEL AVERAGING

Bayesian neural networks impose significant challenge due to computational complexity both during training and inference. Bayesian Model Averaging (Wasserman, 2000) is often applied in practice. This entails approximating the predictive distribution by drawing M samples of the parameter vector θ_i from the posterior distribution and averaging their corresponding predictions p_i :

$$\tilde{p} = \frac{1}{M} \sum_{i=1}^M p_i = \frac{1}{M} \sum_{i=1}^M h_{\theta_i}^{\text{bnn}}(x). \quad (12)$$

In practice, these M predictive samples may arise either from repeated posterior sampling of a single Bayesian Neural Network or from an ensemble of M independently trained models. The ensemble interpretation is particularly attractive in large-scale applications, as it offers a straightforward, parallelizable, and empirically effective approach to approximate Bayesian Model Averaging while providing robust uncertainty estimates.

E UNCERTAINTY DECOMPOSITION IN ENSEMBLE AND BAYESIAN MODELS

Given M models sampled with Bayesian Model Averaging or trained independently in an ensemble, the uncertainty can be decomposed using Shannon entropy (Hüllermeier & Waegeman, 2021). The *total uncertainty* TU is the entropy of the final averaged prediction, while the *aleatoric uncertainty* AU is the average entropy of the individual sample predictions:

$$\text{TU} := H(\tilde{\mathbf{p}}) = - \sum_{k=1}^C \tilde{p}_k \log_2 \tilde{p}_k, \quad (13)$$

$$\text{AU} := \tilde{H}(\mathbf{p}) = \frac{1}{M} \sum_{i=1}^M H(\mathbf{p}_i) = - \frac{1}{M} \sum_{i=1}^M \sum_{k=1}^C p_{i_k} \log_2 p_{i_k}, \quad (14)$$

where C is the number of classes. The *epistemic uncertainty* EU is then the difference, $\text{EU} := \text{TU} - \text{AU}$, which quantifies the disagreement among the posterior samples (Hüllermeier et al., 2022).