Random-Set Graph Neural Networks

Tommy Woodley[*]
School of Engineering,
Computing and Mathematics
Oxford Brookes University
19204799@brookes.ac.uk

Matteo Tolloso[*]
Department of Computer Science
University of Pisa
matteo.tolloso@phd.unipi.it

Davide Bacciu
Department of Computer Science
University of Pisa
davide.bacciu@unipi.it

Shireen Kudukkil Manchingal
Oxford Dynamics
shireen.manchingal@oxdynamics.com

Fabio Cuzzolin
Oxford Brookes Institute for
Artificial Intelligence,
Data Analysis and Systems (AIDAS)
fabio.cuzzolin@brookes.ac.uk

Abstract

Show a graph neural network a node from a class it has never seen, and it will confidently name one of the classes it knows. In safety-critical settings, where a node's neighborhood can be novel, conflicting, or rare in the training graph, this is a serious flaw. The culprit is the softmax head, which must spread all of its confidence over the known classes and so has no way to say "I don't know". We introduce *Random-Set Graph Neural Networks (RS-GNN)*, a drop-in replacement for the softmax head of any message-passing GNN. Instead of one probability per class, RS-GNN assigns belief to *sets* of classes, so it can express statements such as "this node is one of these two classes, but the graph does not tell me which". Each prediction thus becomes a whole set of plausible class distributions: its center gives the predicted label, and its width measures what the model does not know. To keep this tractable, RS-GNN reasons over a small, graph-aware family of class sets, chosen from classes that overlap in the learned embedding space. On nine node-classification benchmarks, both homophilic and heterophilic, RS-GNN matches or outperforms probabilistic, ensemble, and credal baselines at detecting nodes from unseen classes. On road-scene graphs built from ROAD and nuScenes, it substantially improves accuracy on nuScenes and, when moved to a new dataset, degrades more gracefully than a softmax head, which becomes fully overconfident. RS-GNN lets a graph model report not only what it predicts, but also how much it does not know.

1 Introduction

Machine learning systems are increasingly deployed in safety-critical domains where erroneous predictions may lead to severe consequences. In such settings, it is not sufficient for a model to be accurate; it must also be aware of its own limitations: in other words, to know when it does not know.

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

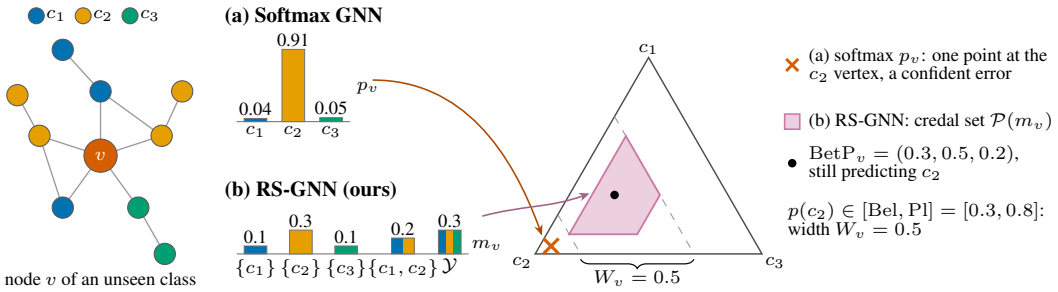


Figure 1: **Softmax vs. random-set head on a node of an unseen class.** Node colours denote the known classes $\mathcal{Y} = \{c_1, c_2, c_3\}$; node v (red) belongs to a class never seen in training. *(a)* A softmax GNN outputs a single distribution p_v, a point ($\times$) at the c_2 vertex of the simplex: a confident error. *(b)* RS-GNN outputs a mass function m_v over focal sets, including non-singleton ones (bars split by the classes they contain), which induces a whole credal set of distributions (shaded). Its pignistic probability BetP_v (•) is the point prediction, and the range of $p(c_2)$ over the credal set (brace) is the width $W_v = 0.5$ that quantifies epistemic uncertainty. Numbers are those of the worked example in Sec. 3.

A central obstacle to trustworthy deployment is the reliable quantification of *epistemic uncertainty*, the uncertainty arising from limited knowledge about the data-generating process [19, 21].

This challenge is particularly acute for Graph Neural Networks (GNNs), which are increasingly deployed in real-world systems where acting on an overconfident prediction is costly. Unlike classical i.i.d. learning, graph data exhibits strong relational dependencies between instances, and predictions must aggregate information over many hops, a regime in which message passing is known to struggle [33]. Uncertainty can therefore originate not only from noisy node features but also from structural ambiguity in the graph topology. A node may be uncertain because its neighborhood is inherently ambiguous (aleatoric uncertainty), or because its structural context is novel and insufficiently represented in the training data (epistemic uncertainty) [10, 21]. Disentangling these sources is essential for robust out-of-distribution (OOD) detection and reliable graph learning [11, 46], in applications ranging from e-commerce recommendation [45] to molecular optimization [32].

Bayesian approaches provide a principled lens for this decomposition by marginalizing predictions over a posterior distribution of model parameters [29]. In GNNs, Bayesian variants place distributions over weights and, in some cases, over the graph structure itself [16, 35]. However, Bayesian Graph Neural Networks often incur substantial computational overhead and require carefully specified priors [12, 47]. Ensemble-based approaches offer a practical alternative by measuring disagreement across independently-trained models [22, 30]; yet, they scale poorly and remain limited to second-order probability representations. Recent work has explored *credal learning* on graphs [44], where models output probability intervals forming a credal set [6, 48]. While credal methods provide expressive set-valued predictions and principled uncertainty decompositions, they operate within the space of probability distributions.

Standard softmax GNNs are constrained to place all predictive mass on the probability simplex (Fig. 1a): a node whose neighbourhood is consistent with several classes still receives a single sharp distribution. The simplex has no place to record "the label belongs to $\{c_1, c_2\}$ rather than to either alone", so structural ambiguity collapses into a confident point estimate, and genuine epistemic uncertainty becomes indistinguishable from a confident error.

In contrast, *random sets* offer a more fundamental representation of epistemic uncertainty. A random set is a set-valued random variable that assigns probability mass directly to subsets of outcomes [34, 39]. On finite domains, random sets take the form of *belief functions*, which generalize classical probability measures and naturally induce a convex set of consistent probability distributions (a credal set) [8]. In Random-Set Neural Networks (RS-NN), predictions are expressed as belief functions over the class list, encoding epistemic uncertainty through the size and geometry of the associated credal set. A point prediction can then be recovered via the pignistic transformation, corresponding geometrically to the center of mass of the credal set.

While RS-NNs perform well in i.i.d. settings such as image classification and OOD detection, their formulation does not transfer to graphs unmodified. Two challenges arise that we address directly. First, message passing violates the i.i.d. assumption by construction, so each node's belief assignment must be conditioned on its message-passing representation rather than predicted from an independent input. Second, heterophilic graphs blur class boundaries because connected nodes may belong to different classes [26, 53]; thus, the family of *focal sets* (sets of classes with non-zero mass) must be graph-aware, rather than uniform over the power set.

In this paper, we introduce *Random-Set Graph Neural Networks (RS-GNNs)*, a random-set based method for epistemic uncertainty quantification on graphs (Fig. 1b). RS-GNN replaces the classical softmax layer of a GNN with a belief-output layer that predicts mass functions over a budgeted collection of focal sets (sets of classes). Each node is thus mapped to a belief function, which is mathematically equivalent to a credal set of probability distributions on the class list. The pignistic probability provides a central prediction, while the geometry of the induced credal set quantifies epistemic uncertainty.

The code to reproduce all the experiments is available at `https://github.com/sami-nii/random-set-graphs`.

The following are the main **contributions** of this paper:

- **Random-Set Graph Neural Networks.** The first *random-set based method* for graph-structured data: a node-wise belief head that places mass on non-singleton focal sets $A \subseteq \mathcal{Y}$ ($|A| > 1$) and reads off the credal-set width as an explicit epistemic-uncertainty score.
- **Graph-aware focal-set construction.** A scalable, budgeted focal-set selection strategy that allocates non-singleton focal sets according to class-overlap statistics in the learned node-embedding space, sidestepping the 2^C blow-up of full power-set enumeration.
- **Empirical evidence across nine standard benchmarks and two temporal road-scene benchmarks.** On leave-out-class OOD detection, RS-GNN reaches an AUROC of 88.84 on Cora and 84.18 on Chameleon, matching or outperforming Classical Ensemble, GEBM, CaGCN, and recent credal baselines; on temporal road-scene graphs (ROAD, nuScenes), RS-GNN improves nuScenes node-classification accuracy from 0.418 to 0.593 and degrades less severely than a softmax head under cross-dataset shift.

2 Related work

Uncertainty quantification in Graph Neural Networks (GNNs) [2] is a key challenge due to relational dependencies absent in i.i.d. settings. Recent surveys highlight rapid progress in predictive reliability and robustness under distribution shift [7, 47]. Existing approaches fall into four families: deterministic methods, Bayesian GNNs, ensembles, and set-valued representations.

Deterministic methods use a single trained model. Post-hoc heuristics like maximum softmax probability or predictive entropy are efficient but poorly calibrated [13]. More expressive approaches include evidential methods predicting higher-order distribution parameters (e.g., Dirichlet) [42, 52], energy-based methods [50], post-hoc calibration mappings [49], and the Graph Energy-Based Model (GEBM) for epistemic uncertainty [9]. These reduce uncertainty to scalar probabilistic outputs.

Bayesian GNNs extend probabilistic inference to graphs via distributions over parameters and graph components. Adaptive connection sampling [16], Monte Carlo Dropout [10], and Variational Inference [17] approximate posteriors, capturing uncertainty in weights, features, and structure [35]. Graph Posterior Networks perform explicit Bayesian updates [42]. However, they incur high computational cost and require careful priors [20]. Ensembles aggregate predictions from multiple trained models [4, 30], offering empirical strength but linear cost in model count. Their uncertainty remains grounded in probability distributions.

Set-valued methods represent uncertainty via sets of plausible distributions. Credal learning uses probability intervals or distribution sets [6, 44, 48] but stays in probability space. Random sets assign mass to subsets of outcomes [34, 39], yielding belief functions and credal sets [8]. Random-Set Neural Networks (RS-NN) predict belief functions for uncertainty estimation in i.i.d. settings [31] but do not address graph dependencies. Unlike evidential GNNs [42, 52], which place mass only on singletons plus a single total-uncertainty term, RS-GNN places mass on *non-singleton* focal sets. This

integrates belief-function representations with message passing, modeling uncertainty as a structured, set-valued quantity that reflects both feature ambiguity and graph-induced dependencies.

3 Methodology

Problem setting. Let $G = (V, E, X)$ be an attributed graph, where V is the set of nodes, E the edge set, and $X \in \mathbb{R}^{|V| \times d}$ the node-feature matrix. The label space is partitioned into in-distribution (ID) classes $\mathcal{Y} = \{1, \dots, C\}$ and out-of-distribution (OOD) classes $\mathcal{Y}_{\text{OOD}}$, with $\mathcal{Y} \cap \mathcal{Y}_{\text{OOD}} = \emptyset$. Labels are observed only for a set of training nodes whose classes lie in $\mathcal{Y}$. For every test node v, the model must output *(i)* a class prediction $\hat{y}_v \in \mathcal{Y}$, evaluated by accuracy and calibration on ID nodes, and *(ii)* an uncertainty score $u_v \in \mathbb{R}$ that should be higher for nodes of $\mathcal{Y}_{\text{OOD}}$ than for ID nodes, evaluated by OOD-detection AUROC. We call this task *uncertainty-aware node classification under distribution shift*.

Terminology. *Aleatoric uncertainty*: irreducible ambiguity in the data (e.g. a neighbourhood that genuinely mixes classes). *Epistemic uncertainty*: uncertainty due to limited knowledge (e.g. a neighbourhood unlike any seen in training), which more data could reduce. *Mass function*: $m : 2^{\mathcal{Y}} \to [0, 1]$ with $m(\emptyset) = 0$ and $\sum_A m(A) = 1$; $m(A)$ is evidence that the label lies in A that cannot be committed to any strict subset of A. *Focal set*: a set A with $m(A) > 0$. *Belief* and *plausibility*: $\text{Bel}(A) = \sum_{B \subseteq A} m(B)$ and $\text{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B)$, the least and the largest probability that A can receive. *Credal set*: $\mathcal{P}(m) = \{p : \ p(A) \geq \text{Bel}(A) \ \forall A \subseteq \mathcal{Y}\}$, all class distributions consistent with m. *Pignistic probability*: $\text{BetP}(c) = \sum_{A \ni c} m(A)/|A|$, the centre of mass of $\mathcal{P}(m)$, used as point prediction. *Credal-set width*: $W = \text{Pl}(\{\hat{y}\}) - \text{Bel}(\{\hat{y}\})$, the length of the probability interval of the predicted class.

Example. For node v of Fig. 1, with $C = 3$, RS-GNN outputs $m(\{c_1\}) = 0.1$, $m(\{c_2\}) = 0.3$, $m(\{c_3\}) = 0.1$, $m(\{c_1, c_2\}) = 0.2$, and $m(\mathcal{Y}) = 0.3$. The pignistic probability is $(0.3, 0.5, 0.2)$, so the prediction is c_2, as for the softmax head. However, the probability of c_2 is only constrained to $[\text{Bel}(\{c_2\}), \text{Pl}(\{c_2\})] = [0.3, 0.8]$, giving $W = 0.5$: half of the mass is evidence that cannot be committed to a single class. The softmax output $(0.04, 0.91, 0.05)$ has no way to express this, which is the distinction between the point prediction and the model's epistemic imprecision that RS-GNN exploits.

We introduce *Random-Set Graph Neural Networks (RS-GNNs)*, a belief-function-based extension of Graph Neural Networks for epistemic uncertainty quantification in node classification. In a standard GNN, message passing produces node embeddings that are mapped to a probability vector by a softmax classifier (Fig. 1a). This forces the model to allocate all predictive mass to singleton classes, even when the available graph evidence is incomplete, ambiguous, or structurally novel.

RS-GNN replaces this probabilistic output layer with a *belief head* that, for each node v, outputs a belief function over a collection of subsets of the class space, conditioned on a message-passing representation of the node and its neighborhood. The belief assigned to v thus reflects both its own features and its neighbourhood, so that epistemic uncertainty can capture structural novelty and neighborhood conflict as well as feature-level ambiguity.

RS-GNN performs belief-function encoding at the *node level* (Fig. 2): each node v is assigned a belief function over a shared focal-set budget $\mathcal{F}$, so that uncertainty concerns the class membership of each individual node, conditioned on its features and graph context. Graph-level or edge-level belief assignments would instead target graph classification or link prediction, which we leave to future work.

A classical probabilistic classifier assigns to each node v a probability vector over $\mathcal{Y}$. In contrast, RS-GNN assigns probability mass to *subsets* of $\mathcal{Y}$. This allows the model to express partial support for composite hypotheses such as "the label belongs to A", where $A \subseteq \mathcal{Y}$, without committing to a single class.

In principle, a belief function over $\mathcal{Y}$ is defined on the full power set $2^{\mathcal{Y}}$, which contains 2^C subsets. This is computationally infeasible when C is large. Following the budgeting principle of RS-NN, RS-GNN therefore operates on a restricted collection of focal sets $\mathcal{F} \subseteq 2^{\mathcal{Y}}$, which always includes all singleton classes $\{\{1\}, \dots, \{C\}\}$ and may include selected non-singleton subsets.

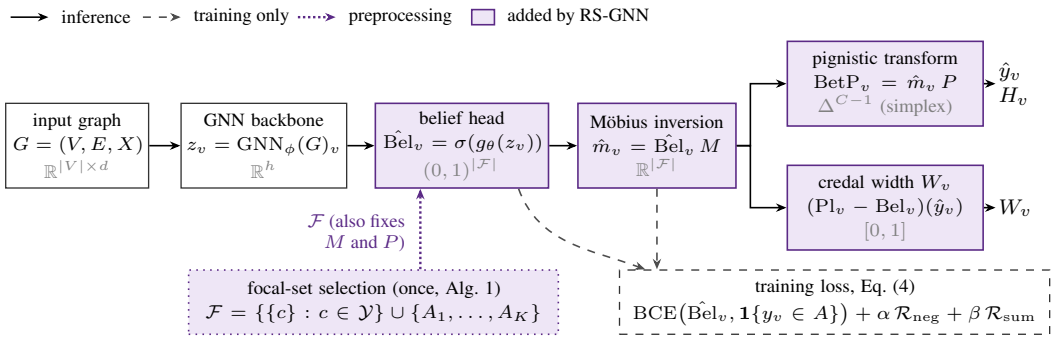


Figure 2: **Overview of the Random-Set Graph Neural Network (RS-GNN).** Shaded blocks are the components RS-GNN adds to a standard message-passing GNN; the grey line in each block gives the dimension of its output. At inference (solid arrows), the belief head predicts beliefs over a budgeted family of focal sets $\mathcal{F}$, the fixed Möbius and pignistic matrices M and P turn them into masses and class probabilities, and the credal interval of the predicted class gives the epistemic score W_v. The focal family is selected once, before training (dotted; Alg. 1), and training (dashed) matches the beliefs to set-valued targets while penalising invalid masses (Eq. 4).

For small class spaces (standard benchmarks with fewer than 10 ID classes), $\mathcal{F}$ is the full non-empty power set. For larger class spaces, we use a budgeted focal-set construction inspired by RS-NN [31], computed once before training from class-overlap statistics in a message-passing node-embedding space: embeddings of a class-balanced sample of labelled ID training nodes are produced by a warm-up GNN with the backbone architecture, trained on labelled ID nodes only, projected to three dimensions with t-SNE, and each class is summarised by the 95% ellipsoid of a Gaussian fitted to its points. Candidate pairs and triples of classes are scored by the Jaccard overlap of their ellipsoids, and the K highest-scoring subsets are added to the singletons (Alg. 1, §D.6). OOD labels are never used. Intuitively, non-singleton focal sets correspond to groups of classes that are difficult to distinguish under message passing, so the budget is graph-aware: rather than allocating capacity uniformly over all possible subsets, RS-GNN concentrates belief mass on class subsets that are empirically relevant for uncertainty representation.

RS-GNN can be built on top of any message-passing GNN. Let $Z^{(0)} = X$ denote the input node features. At layer l, node representations are updated as $Z^{(l)} = \Phi^{(l)}(Z^{(l-1)}, \Psi^{(l)}(Z^{(l-1)}, E))$, where $\Psi^{(l)}$ denotes neighborhood aggregation and $\Phi^{(l)}$ the node-update function. After L layers, each node v has an embedding $z_v = Z_v^{(L)}$, which summarises both node-level features and structural context. A standard GNN would map z_v to logits followed by a softmax distribution over classes. RS-GNN instead maps z_v to belief values over focal sets, $\hat{\mathrm{Bel}}_v(A_k) = \sigma(g_\theta(z_v)_k)$ for $A_k \in \mathcal{F}$, where g_θ is the belief head and σ is the sigmoid function. The sigmoid output is used because belief supervision is set-valued: for a labeled node, multiple focal sets may be consistent with the true class.

A belief function can equivalently be represented by a mass function. Given predicted belief values, RS-GNN obtains mass values through *Möbius inversion*:

$$\hat{m}_v(A) = \sum\nolimits_{B \in \mathcal{F},\, B \subseteq A} (-1)^{|A \setminus B|}\, \hat{\mathrm{Bel}}_v(B), \qquad A \in \mathcal{F}. \tag{1}$$

When $\mathcal{F} = 2^{\mathcal{Y}} \setminus \{\emptyset\}$, Eq. (1) is the exact Möbius inverse on the full power set. When a budgeted focal family is used, the inversion is computed on the selected focal-set lattice induced by $\mathcal{F}$, yielding an approximate belief-to-mass transformation over the retained focal sets. As in RS-NN, the validity of the resulting mass assignment is encouraged through regularization.

A valid mass function should satisfy $\hat{m}_v(A) \geq 0$ and $\sum_{A \in \mathcal{F}} \hat{m}_v(A) = 1$. Because the belief head is trained with soft constraints rather than by explicitly projecting onto the belief-function polytope, small violations may occur during training; we penalise such violations in the objective described below.

For decision making, RS-GNN converts the predicted mass function into a pignistic probability distribution. The pignistic transformation redistributes the mass of each focal set uniformly among its elements, $\widetilde{\mathrm{BetP}}_v(i) = \sum_{A_k \in \mathcal{F}:\, i \in A_k} \hat{m}_v(A_k)/|A_k|$. In practice, because mass validity is only

encouraged during training, negative masses are first clipped to zero, any residual mass is assigned to $\mathcal{Y}$ when $\mathcal{Y} \in \mathcal{F}$, and the mass vector is renormalised before the transformation is applied (Fig. 11), which yields a valid probability distribution BetP_v; the final point prediction is $\hat{y}_v = \arg\max_{i \in \mathcal{Y}} \mathrm{BetP}_v(i)$. Thus, RS-GNN retains a standard class prediction interface while internally representing uncertainty through belief functions.

RS-GNN provides two complementary uncertainty scores corresponding to the standard aleatoric/epistemic decomposition. First, *total predictive* uncertainty (aleatoric + epistemic) is measured by the entropy of the pignistic probability,

$$H_v = -\sum_{i=1}^{C} \mathrm{BetP}_v(i) \log\big(\mathrm{BetP}_v(i) + \varepsilon\big). \tag{2}$$

This quantity is analogous to predictive entropy in probabilistic models, but is computed after projecting a belief function to its pignistic center. Second, the *epistemic* component is isolated through the width of the induced credal set, which depends only on the spread of probability distributions consistent with the predicted mass function and is invariant to aleatoric noise within a single distribution: for each singleton class i, the lower and upper probabilities induced by the mass function are $\underline{P}_v(i) = \hat{m}_v(\{i\})$ and $\overline{P}_v(i) = \sum_{A_k \in \mathcal{F}:\, i \in A_k} \hat{m}_v(A_k)$. The width of the interval $[\underline{P}_v(i), \overline{P}_v(i)]$ measures how much probability can be assigned to class i by distributions compatible with the predicted belief function. Following RS-NN, we use the probability interval width for the predicted class as an epistemic uncertainty score:

$$W_v = \overline{P}_v(\hat{y}_v) - \underline{P}_v(\hat{y}_v). \tag{3}$$

A large value of W_v indicates that there is not enough evidence for the model to commit to a precise singleton class, reflecting epistemic ambiguity in the node prediction.

What is guaranteed. During training, mass validity is only encouraged (see below); at inference, negative masses are clipped and the mass vector is renormalised before computing BetP_v and W_v (Fig. 11), so the following properties hold for every prediction RS-GNN outputs (proof in §D.5).

Proposition 1. *Let m be a valid mass function on $\mathcal{F}$ ($m \geq 0$, $\sum_{A \in \mathcal{F}} m(A) = 1$, $\emptyset \notin \mathcal{F}$). Then* (i) *the credal set $\mathcal{P}(m)$ is non-empty, closed, and convex;* (ii) $\mathrm{BetP} \in \mathcal{P}(m)$*, so the point prediction is consistent with the represented uncertainty;* (iii) $W = \sum_{A \in \mathcal{F}:\, \hat{y} \in A,\, |A| > 1} m(A) \in [0, 1]$*; and* (iv) $\mathrm{Pl}(\{c\}) - \mathrm{Bel}(\{c\}) = 0$ *for every class c if and only if m is supported on singletons, in which case $\mathcal{P}(m) = \{\mathrm{BetP}\}$ reduces to a single (softmax-like) distribution.*

Property *(iv)* makes precise why a softmax head, which is the special case of a mass function supported on singletons, cannot express epistemic imprecision.

Training is performed directly in belief space. For a labeled node v with ground-truth class y_v, the target belief value for a focal set $A \in \mathcal{F}$ is $\mathrm{Bel}_v^{\mathrm{target}}(A) = \mathbf{1}\{y_v \in A\}$: all focal sets containing the true label are assigned target belief 1, and all focal sets not containing the true label are assigned target belief 0. This resembles a multi-label target vector, but its semantics are different: the positive entries do not indicate multiple labels, but rather all propositions that are logically consistent with the observed class. The primary loss is binary cross-entropy over focal sets averaged across labeled training nodes $\mathcal{V}_{\mathrm{tr}}$ and focal sets, $\mathcal{L}_{\mathrm{Bel}} = \mathrm{BCE}(\mathrm{Bel}^{\mathrm{target}}, \hat{\mathrm{Bel}})$. To encourage the predicted belief values to induce valid mass functions, we add two regularisation terms: $\mathcal{R}_{\mathrm{neg}} = \mathbb{E}_v[\max(0, -\hat{m}_v)]$ penalises negative masses, and $\mathcal{R}_{\mathrm{sum}} = \mathbb{E}_v|\sum_{A_k \in \mathcal{F}} \hat{m}_v(A_k) - 1|$ penalises deviations from unit total mass. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{Bel}} + \alpha\, \mathcal{R}_{\mathrm{neg}} + \beta\, \mathcal{R}_{\mathrm{sum}}, \tag{4}$$

where α and β control the strength of the mass-validity constraints.

The additional cost of RS-GNN arises from replacing the C-dimensional softmax output with a $|\mathcal{F}|$-dimensional belief head. Since $\mathcal{F}$ contains all singleton classes and only a budgeted number of non-singleton focal sets, the output dimension grows as $C + K$, rather than 2^C. The method therefore avoids the exponential complexity of full random-set prediction while preserving the ability to represent class-level ambiguity through selected composite focal sets. At inference time, RS-GNN requires only a single forward pass, followed by belief-to-mass conversion and pignistic projection,

making it compatible with standard mini-batch graph training and evaluation. The one-off focal-set construction scores at most $O(C^2 + C^3)$ candidate pairs and triples, and its measured cost is essentially independent of K (Tab. 7). Empirically, the belief head adds 0.07–2.7 ms per forward pass over the backbone, a lower mean overhead than Energy, Mahalanobis, ODIN, and KNN on every dataset (Tab. 9, §D.8).

4 Experiments

We answer three questions, each with a single quantitative measure. **(i)** *OOD detection* (Sec. 4.2): is RS-GNN's uncertainty higher on nodes of unseen classes than on known ones, compared with existing methods (AUROC on nine standard benchmarks)? **(ii)** *Effect of the head* (Sec. 4.3): with the backbone fixed, does replacing softmax with the belief head improve accuracy, calibration, and OOD separation on road-scene graphs (accuracy, ECE, AUROC)? **(iii)** *Cross-dataset shift* (Sec. 4.3): when test graphs come from a different dataset, does the belief head stay less overconfident than softmax (AUROC)? Sec. 4.4 ablates the design choices and reports the computational cost.

4.1 Experimental setup

Datasets. We evaluate RS-GNN on two complementary families of node-classification benchmarks. First, nine *standard graph datasets* spanning homophilic and heterophilic regimes: `Coauthor` [40], `Cora` [51], and `Reddit2` [15] are homophilic; `Chameleon`, `Squirrel` [38], `Roman Empire`, `Amazon-Ratings` [37], and the year-prediction citation graphs `ArXiv` [18, 28] and `Patents` [24, 26, 28] are heterophilic (statistics in Tab. 10). Second, two *temporal road-scene benchmarks* constructed from autonomous-driving data: `ROAD` [41] (49,123 train / 26,302 val / 26,301 test graphs; 291,197 labeled test nodes) and `nuScenes` [5] (307 / 39 / 38 graphs). Each temporal window is encoded as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$ where nodes are scene context and traffic entities, edges encode spatial-temporal relations, and the task is node classification. A graph is natural here: an agent's category is ambiguous from its own box and action attributes alone (e.g. under occlusion or at long range), but becomes identifiable from the scene context and from the same agent's track in adjacent frames, which message passing aggregates. Full dataset descriptions, temporal-graph construction details, and feature schemas are in §A.

Leave-out-class OOD protocol. We use a transductive leave-out-class setup: the class set is partitioned into in-distribution classes $\mathcal{Y}$ and out-of-distribution classes $\mathcal{Y}_{\text{OOD}}$, with $\mathcal{Y} \cap \mathcal{Y}_{\text{OOD}} = \emptyset$. During training, OOD-class labels are masked; at test time, labels may belong to either set. The objective is the one defined in Sec. 3: classify ID nodes correctly while assigning high uncertainty to OOD nodes. OOD nodes remain in the graph and take part in message passing, but never in the loss; this is applied identically to all methods (§D.9), so the standard-benchmark results measure *transductive* leave-out-class OOD detection. The cross-dataset setting below complements it with a stronger, inductive shift, since nuScenes graphs are never seen during training. For road-scene experiments we use $\mathcal{Y}_{\text{OOD}} = \{8, 9\}$ on ROAD and $\mathcal{Y}_{\text{OOD}} = \{6, 7\}$ on nuScenes; we additionally evaluate a cross-dataset setting in which ROAD classes are treated as ID and nuScenes test nodes as OOD at evaluation time.

Baselines. We compare against established uncertainty-estimation and OOD-detection baselines [28]: post-hoc methods (`Energy` [27], `ODIN` [25], `Mahalanobis` [23], `KNN` [43], `CaGCN` [49]); graph-specific methods (`GNNSafe` [50], `GEBM` [9]); credal learning methods (`Credal final`, `Credal ensemble`, `Credal frozen`, `CredalLJ`) [44]; `JLDE`; and the `Classical Ensemble` [1, 14, 22, 36], the de-facto reference for uncertainty estimation. All models share the same GATv2 backbone [3] (§D.2).

4.2 OOD detection on standard graph benchmarks

Tab. 1 reports leave-out-class OOD AUROC for all methods. No single method dominates all nine benchmarks, so we first look at aggregate performance: RS-GNN obtains the most first places on the primary score (Chameleon 84.18, ArXiv 70.21, Roman Empire 69.58), is within 0.01 of the best method on Cora (88.84 vs. 88.85 for GNNSafe), and has the second-best mean AUROC (69.98) and average rank (4.44), behind only the post-hoc calibration method CaGCN (70.48, 4.33) and

Table 1: OOD detection performance (AUROC ↑, %) on the nine standard benchmarks. For methods that expose two uncertainty scores, both AUROCs are reported as *total/epistemic* (for RS-GNN: pignistic entropy / credal-set width). In each dataset column, the best value is in **bold**, the second best is underlined, and the third best is in *italics*, ranking all reported values. *Mean* and *Rank* are the mean AUROC and the average rank across datasets, computed on the first value of each cell (lower rank is better).

	Heterophilic						Homophilic			Aggregate	
Method	Chameleon	Squirrel	ArXiv	Patents	Amazon-Ratings	Roman Empire	Coauthor	Reddit2	Cora	Mean ↑	Rank ↓
Post-hoc, graph-specific, and ensemble baselines											
Energy	58.25	44.47	50.16	43.33	51.42	52.26	94.47	43.93	12.59	50.10	9.44
KNN	56.81	52.79	56.86	52.55	52.46	29.32	88.23	65.63	84.67	59.92	7.78
ODIN	57.98	47.28	48.16	43.21	46.23	53.98	94.17	43.09	71.65	56.19	10.22
Mahalanobis	51.82	53.79	59.52	58.72	49.66	67.26	82.49	68.98	80.26	63.61	7.00
GNNSafe	50.42	35.88	35.30	27.35	49.50	50.28	*94.82*	61.99	**88.85**	54.93	9.33
Classical ensemble	74.00/30.22	58.32/59.13	58.20/65.45	48.23/60.35	48.51/51.36	46.49/55.39	95.30/94.81	58.84/*72.09*	75.85/43.84	62.64	7.67
JLDE	70.06	71.06	45.35	46.74	51.72	31.01	32.37	36.55	55.36	48.91	10.11
GEBM	45.41	43.00	56.26	54.49	48.79	62.62	92.20	54.49	87.15	60.49	8.67
CaGCN	73.77	60.05	58.30	54.15	51.52	68.03	**96.38**	**91.13**	81.03	**70.48**	**4.33**
Credal methods											
Credal final	76.29/67.27	75.02/65.85	64.65/65.66	68.92/69.73	51.89/48.11	31.51/68.48	74.80/50.74	70.57/69.35	43.79/56.17	61.94	5.67
Credal ensemble	74.98/29.54	59.03/53.66	58.05/50.97	47.41/64.64	48.59/51.17	45.90/53.06	93.85/93.72	61.16/57.72	75.17/36.42	62.68	8.11
Credal frozen	48.97/**85.67**	73.03/26.96	42.36/**70.22**	33.29/*68.97*	*52.41*/48.78	68.05/**69.75**	93.38/43.78	64.69/73.63	84.97/17.33	62.35	7.22
CredalLJ	72.37/77.67	*73.89*/**77.04**	65.77/63.79	**70.78**/60.06	**52.82**/47.27	31.45/65.68	86.58/70.88	66.35/67.31	83.26/21.95	*67.03*	*5.00*
RS-GNN (ours)	84.18/*84.10*	63.57/63.56	70.21/*69.39*	49.32/42.14	51.05/50.81	69.58/*69.39*	85.72/74.56	67.38/66.85	88.84/*88.51*	69.98	4.44

ahead of all credal and ensemble methods. On **Reddit2** and **Coauthor**, dedicated calibration-style baselines (CaGCN, GNNSafe) retain the lead, while on **Patents**, **Squirrel**, and **Amazon-Ratings** the credal baselines (CredalLJ, Credal final) are stronger. RS-GNN's two scores are also consistent: its entropy- and credal-width AUROCs differ by less than one point on seven of nine datasets, whereas for credal and ensemble baselines the two scores can differ by more than 35 points (e.g. Credal frozen on Chameleon, 48.97 vs. 85.67, and on Cora, 84.97 vs. 17.33), so which of their scores to deploy cannot be decided without OOD data.

4.3 Temporal road-scene graphs

In-distribution performance. Tab. 2 reports ID metrics on the road-scene benchmarks. On **ROAD held-out OOD**, both models achieve comparable accuracy (0.754 vs. 0.746), but RS-GNN improves calibration substantially (**ECE 0.132 vs. 0.178**; lower NLL and Brier), indicating better probabilistic reliability without sacrificing accuracy. On **nuScenes** the difference is more pronounced: RS-GNN **improves accuracy from 0.418 to 0.593** and substantially **reduces NLL (7.047 to 2.079)**. This suggests that belief-based modeling stabilises training under stronger class imbalance and shorter temporal windows. In the **cross-dataset ROAD(ID)→nuScenes(OOD)** setting, RS-GNN slightly improves ID accuracy (0.770 vs. 0.751), although vanilla softmax achieves better ECE and NLL on ROAD alone. This reflects the expected trade-off: *random-set heads prioritise uncertainty separation over pure closed-set likelihood optimisation.*

Table 2: In-distribution (ID) performance across ROAD and nuScenes settings. Metrics are reported on ID nodes only. For ROAD(ID)-nuScenes(OOD), ID metrics are computed on ROAD test nodes.

Setting	Dataset	Split	Model	Uncertainty	Acc (↑)	ECE (↓)	NLL (↓)	Brier (↓)
ROAD held-out OOD	ROAD	ID: $\{0,\dots,7\}$	Vanilla	Softmax entropy	**0.754**	0.178	1.849	0.431
		OOD: $\{8,9\}$	RS-GNN	Pignistic entropy	0.746	**0.132**	**1.478**	**0.384**
nuScenes held-out OOD	nuScenes	ID: $\{0,\dots,5\}$	Vanilla	Softmax entropy	0.418	0.452	7.047	0.905
		OOD: $\{6,7\}$	RS-GNN	Pignistic entropy	**0.593**	**0.422**	**2.079**	**0.875**
ROAD(ID) vs nuScenes(OOD)	ROAD	ID: all ROAD classes	Vanilla	Softmax entropy	0.751	**0.122**	**1.035**	**0.381**
		OOD: nuScenes	RS-GNN	Pignistic entropy	**0.770**	0.157	1.837	0.398

OOD uncertainty separation. Tab. 3 (visualised in Fig. 9, §B) reports entropy and credal-width behaviour. On **ROAD held-out OOD**, vanilla entropy separation is weak (AUROC 0.425), while RS-GNN **substantially improves entropy-based OOD detection (AUROC 0.662)**; its credal width, however, is lower on OOD than on ID nodes (AUROC 0.342), so on ROAD the pignistic entropy is the informative score. On **nuScenes held-out OOD**, vanilla softmax entropy is almost zero for both ID and OOD nodes (1.2×10^{-4} vs. 9.8×10^{-5}), i.e. the model is uniformly overconfident, and separation is near chance (AUROC 0.463); RS-GNN is also close to chance (entropy AUROC 0.500; credal-width AUROC 0.521), so on this small benchmark its benefit is in ID accuracy rather than

OOD detection. Under the **ROAD→nuScenes cross-dataset shift**, both models are *more* confident on foreign-domain nodes than on ID nodes (AUROC well below 0.5). The softmax head is the extreme case: its entropy on nuScenes nodes is numerically zero (2.5×10^{-10}) and its AUROC is 0.001, i.e. every foreign node is more confidently classified than every ID node. RS-GNN is less extreme (entropy AUROC 0.053; credal-width AUROC 0.083), but neither model detects this shift reliably. Fig. 4 in the appendix illustrates the two heads on a road-scene graph with an agent of an unseen type.

Table 3: Out-of-distribution uncertainty and detection evaluation across ROAD and nuScenes settings. For Vanilla, uncertainty is softmax entropy; for RS-GNN, uncertainty is pignistic entropy with credal-width metrics where available. Bold marks the better model where both models report the metric; AUROC below 0.5 means that OOD nodes receive *lower* uncertainty than ID nodes.

Setting	Model	In-distribution (ID)		Out-of-distribution (OOD)		OOD Detection Metrics			
		Entropy (ID) (↓)	Credal (ID) (↓)	Entropy (OOD) (↑)	Credal (OOD) (↑)	AUROC (Entropy) (↑)	AUPRC (Entropy) (↑)	AUROC (Credal) (↑)	AUPRC (Credal) (↑)
ROAD held-out OOD	Vanilla	0.180	–	0.226	–	0.425	0.208	–	–
($\mathcal{Y}_{OOD} = \{8, 9\}$)	RS-GNN	**0.158**	0.094	**0.354**	0.017	**0.662**	**0.327**	0.342	0.186
nuScenes held-out OOD	Vanilla	**0.000124**	–	0.000098	–	0.463	0.071	–	–
($\mathcal{Y}_{OOD} = \{6, 7\}$)	RS-GNN	0.045	2.057	**2.079**	2.061	**0.500**	**0.077**	0.521	0.076
ROAD (ID)	Vanilla	0.329	–	2.48×10^{-10}	–	0.001	0.470	–	–
vs nuScenes (OOD)	RS-GNN	**0.200**	0.065	**1.86**$\times 10^{-7}$	0.121	**0.053**	**0.500**	0.083	0.500

4.4 Ablation and computational cost

Which ingredients matter. We ablate the focal-set construction and the loss on Cora (full power set of the ID classes) and Roman Empire (budgeted focal sets). Both the belief-space objective and the non-singleton focal sets contribute (Tab. 4): the full model reaches the best AUROC on both datasets, removing the non-singleton focal sets costs 2–6 points, and replacing the belief-space objective with plain BCE, with or without margin regularisation, costs up to 16 points. Singletons-only runs have a slightly higher mean AUROC across the sweep but a larger run-to-run variance (§C). Modelling a few confusable class subsets therefore suffices, at a modest increase in output dimensionality.

Table 4: Ablation: best OOD AUROC over the sweep (details in §C).

Configuration	Cora	Roman E.
full RS-GNN	**0.731**	**0.703**
singletons only	0.674	0.681
BCE only	0.642	0.572
BCE + margin	0.611	0.545

Computational cost. The belief head adds 0.07–2.7 ms per forward pass over the backbone, a lower mean overhead than Energy and Mahalanobis on every dataset and roughly 80 to 6,000 times less than ODIN and KNN (Tab. 9). The one-off focal-set construction takes 22–43 s on Coauthor, Reddit2, and Roman Empire, essentially independently of the budget $K \in \{4, \ldots, 64\}$ (Tab. 7), and is never repeated at inference.

5 Discussion and limitations

When does set-valued uncertainty help? RS-GNN's first places are all on heterophilic graphs (Chameleon, ArXiv, Roman Empire), where connected nodes often belong to different classes and neighbourhood evidence is conflicting, and it is also near-best on the homophilic Cora. By contrast, on the homophilic Reddit2 and Coauthor, where the backbone already produces a well-ranked probabilistic output, calibration-focused baselines remain stronger. Heterophily alone is not sufficient: on Squirrel, Patents, and Amazon-Ratings the credal baselines lead. Squirrel is about $2.8\times$ denser than Chameleon (38 vs. 14 edges per node), so aggregation mixes many more conflicting neighbours; we hypothesise that this blurs the class-overlap statistics from which focal sets are selected, but have not verified it. Feature sparsity does not explain the pattern either (§D.10): RS-GNN ranks first both on a very sparse dataset (Chameleon, 99.4% zero features) and on dense ones (ArXiv, Roman Empire), and ranks low on both sparse (Coauthor, Patents) and dense (Amazon-Ratings) datasets.

What happens under cross-dataset shift. A softmax head must place all mass on singleton classes, so under shift it can only become confidently wrong: on ROAD→nuScenes its entropy on foreign

nodes is numerically zero. RS-GNN can move mass to non-singleton focal sets and partly does so (its mean OOD credal width is 0.121 rather than zero), but not enough to rank foreign nodes above ID ones. RS-GNN therefore degrades less severely than softmax, but detecting cross-dataset shift remains an open problem for both models.

Limitations. *(i) Calibration-bottlenecked datasets.* On Reddit2 and Coauthor, dedicated post-hoc calibration baselines (CaGCN, GNNSafe) retain the lead over RS-GNN; in those settings, the bottleneck is calibration of an already-good probabilistic ranking rather than representational expressiveness. Adding post-hoc isotonic calibration on top of RS-GNN improves ID ECE on seven of nine datasets, but its effect on OOD AUROC is dataset-dependent (it helps on four datasets and hurts on five, including Coauthor; Tab. 12), so it is a complementary stage rather than a remedy. *(ii) Focal-set budget design.* The focal sets are selected in a fixed preprocessing step rather than learned end to end, and the budget K is a hyperparameter. *(iii) Extreme cross-dataset shift.* Under ROAD→nuScenes, both vanilla softmax and RS-GNN rank foreign nodes as *less* uncertain than ID nodes (entropy AUROC 0.001 and 0.053; credal-width AUROC 0.083); RS-GNN degrades less severely, but it does not solve cross-dataset OOD detection. *(iv) Configuration sensitivity.* On Patents and Reddit2, RS-GNN is sensitive to focal-set construction and seed. *(v) Protocol and validity.* Our standard-benchmark results are transductive: we do not evaluate a strictly inductive leave-out-class variant in which OOD nodes and their edges are removed during training. Mass validity is exact at inference but only soft during training (see §D.4 for a hard-valid alternative).

6 Conclusion

We presented *Random-Set Graph Neural Networks*, a random-set based method that replaces the softmax head of any message-passing backbone with a node-wise belief function over a budgeted family of focal sets at a negligible additional forward-pass cost. The framework recovers an epistemic uncertainty signal (credal-set width) that scalar softmax-derived scores cannot, with the largest gains on heterophilic graphs, and it degrades less severely than softmax under cross-dataset shift. Future directions include learning the focal-set family end to end, calibration stages that also preserve OOD ranking, and graph- and edge-level extensions.

Broader impact. RS-GNN aims to make graph-based predictions safer by flagging nodes whose neighbourhood is novel or ambiguous, which can reduce overconfident errors in safety-critical applications such as road-scene understanding for automated driving. The main risk is over-reliance: its uncertainty estimates are not guaranteed to detect every distribution shift (Sec. 5, in particular cross-dataset shift), so they should complement rather than replace other safeguards. All experiments use publicly available benchmarks and involve no human subjects; for ROAD and nuScenes we use only the released object-level annotations, not raw images, and we do not attempt to identify individuals.

Acknowledgments and Disclosure of Funding

Shireen Kudukkil Manchingal was funded by the Epistemic AI project (E-pi), under the European Union's Horizon 2020 Research and Innovation programme, Grant Agreement No. 964505. The authors declare no competing interests.

References

[1] 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.

[2] Davide Bacciu, Federico Errica, Alessio Micheli, and Marco Podda. A gentle introduction to deep learning for graphs. *Neural Networks*, 129:203–221, 2020.

[3] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks?, 2022. URL `https://arxiv.org/abs/2105.14491`.

[4] 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.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. URL `https://arxiv.org/abs/1903.11027`.

[6] Michele Caprio, Maryam Sultana, Eleni Elia, and Fabio Cuzzolin. Credal learning theory. *Advances in Neural Information Processing Systems*, 37:38665–38694, 2024.

[7] 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.

[8] Fabio Cuzzolin. Uncertainty measures: A critical survey. *Information Fusion*, page 102609, 2024.

[9] Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. Energy-based epistemic uncertainty for graph neural networks. *Advances in Neural Information Processing Systems*, 37: 34378–34428, 2024.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[11] 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.

[12] 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*, pages 45–87. Springer, 2020.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[14] 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*, pages 318–319, 2020.

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[16] 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*, pages 4094–4104. PMLR, 2020.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[23] 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.

[24] 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*, pages 177–187, 2005.

[25] 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.

[26] 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.

[27] 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.

[28] Longfei Ma, Yiyou Sun, Kaize Ding, Zemin Liu, and Fei Wu. Revisiting score propagation in graph out-of-distribution detection. 37:4341–4373, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/08342dc6ab69f23167b4123086ad4d38-Abstract-Conference.html`.

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[30] 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.

[31] Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set neural networks (rs-nn), 2025. URL `https://arxiv.org/abs/2307.05772`.

[32] Luca Miglior, Lorenzo Simone, Marco Podda, and Davide Bacciu. Towards efficient molecular property optimization with graph energy based models. In *ESANN 2025 proceesdings*, ESANN 2025, page 289–294. Ciaco - i6doc.com, 2025. doi: 10.14428/esann/2025.es2025-120. URL `http://dx.doi.org/10.14428/esann/2025.ES2025-120`.

[33] Luca Miglior, Matteo Tolloso, Alessio Gravina, and Davide Bacciu. Can you hear me now? a benchmark for long-range graph propagation. In *International Conference on Learning Representations*, volume 2026, pages 125144–125178, 2026.

[34] Ilya S Molchanov. *Theory of random sets*, volume 19. Springer, 2005.

[35] 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.

[36] 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.

[37] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

[38] Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021.

[39] Glenn Shafer. A theory of statistical evidence. In W. L. Harper and C. A. Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, volume 2, pages 365–436. Reidel, Dordrecht, 1976.

[40] Oleksandr Shchur, Maximilian Mumme, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

[41] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, Jordan Omokeowa, Stanislao Grazioso, Andrew Bradley, Giuseppe Di Gironimo, and Fabio Cuzzolin. Road: The road event awareness dataset for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1036–1054, January 2023. ISSN 1939-3539. doi: 10.1109/tpami.2022.3150906. URL `http://dx.doi.org/10.1109/TPAMI.2022.3150906`.

[42] 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, pages 18033–18048. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/hash/95b431e51fc53692913da5263c214162-Abstract.html`.

[43] Yiyou 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.

[44] Matteo Tolloso and Davide Bacciu. Credal graph neural networks. *arXiv preprint arXiv:2512.02722*, 2025.

[45] Matteo Tolloso, Davide Bacciu, Shahab Mokarizadeh, and Marco Varesi. Real-time and personalized product recommendations for large e-commerce platforms. In *International Conference on Artificial Neural Networks*, pages 279–293. Springer, 2025.

[46] 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)*, pages 1508–1516. IEEE, 2022.

[47] 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.

[48] 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, 2024.

[49] Xiao Wang, Hongrui Liu, Chuan Shi, and Cheng Yang. Be confident! towards trustworthy graph neural networks via confidence calibration. *Advances in Neural Information Processing Systems*, 34:23768–23779, 2021.

[50] 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.

[51] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48. PMLR, 2016.

[52] 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., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/968c9b4f09cbb7d7925f38aea3484111-Abstract.html`.

[53] 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 Datasets

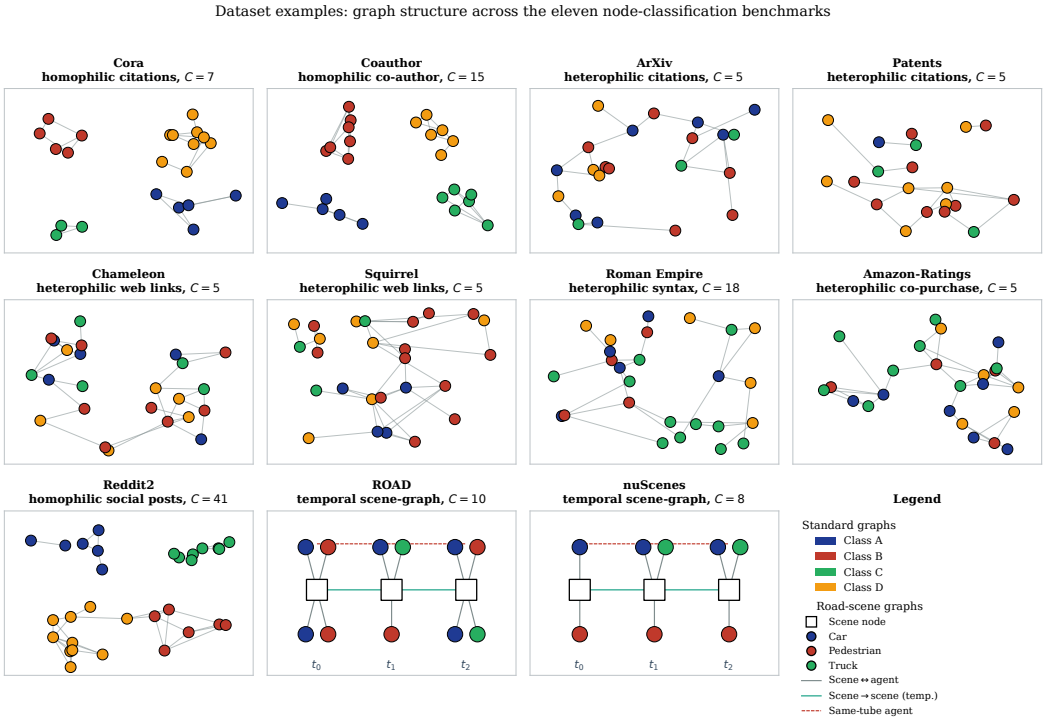


Figure 3: Illustrative graph structure for each of the eleven benchmarks. The toy graphs are schematic: in homophilic panels same-class nodes cluster together, in heterophilic panels they are dispersed. **Top row:** citation and co-authorship networks (Cora and Coauthor are homophilic; the year-prediction citation graphs ArXiv and Patents are heterophilic). **Middle row:** heterophilic web-link, syntax, and co-purchase networks. **Bottom row:** the homophilic social-post benchmark Reddit2 and the two temporal road-scene benchmarks ROAD and nuScenes, which are constructed as scene-graph sequences with intra-frame scene-centred edges (grey), temporal scene-to-scene edges (teal), and same-tube agent edges (red dashed).

Figure 3 shows the structural character of each benchmark. The distinction between homophilic and heterophilic regimes is one of the factors that shape whether scalar softmax-derived uncertainty signals suffice or set-valued representations help (Sec. 5); the temporal scene-graph construction for ROAD and nuScenes is described below.

A.1 Standard graph benchmarks

We evaluate on a diverse suite of standard graph datasets spanning homophilic and heterophilic regimes, as well as different domains and graph scales.

The `Coauthor` dataset is a computer science co-authorship network where nodes represent authors and edges indicate co-authorship relations [40]. 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 hyperlink networks where nodes correspond to web pages and edges correspond to hyperlinks [38]. Node features are derived from informative nouns appearing on each page, and the classification task is to predict page categories based on average monthly traffic. Although these datasets have been criticized for train-test leakage caused by duplicate nodes [37], they remain widely used in the graph-learning literature and are therefore included for comparability.

The Reddit2 dataset is constructed from Reddit posts, where nodes represent posts and edges connect posts if the same user has commented on both [15]. Node features are generated from text embeddings, and the task is to predict the subreddit to which each post belongs.

The ArXiv dataset is a citation network where nodes represent academic papers and directed edges represent citation links [18, 28]. Node features are obtained from embeddings of paper titles and abstracts, and the prediction task is to classify papers according to publication year.

The Roman Empire dataset is based on the English Wikipedia article about the Roman Empire [37]. Each node corresponds to a non-unique word occurrence in the text, and edges connect words that either appear consecutively or are linked in the dependency tree of a sentence. Node labels correspond to syntactic roles, namely the 17 most frequent roles plus an additional class grouping all other roles. Node features are derived from word embeddings.

The Amazon-Ratings dataset is derived from the Amazon product co-purchasing network [37]. Nodes represent products, and edges connect products that are frequently bought together. The task is to predict the average product rating, discretized into five classes. Node features are computed as the mean of word embeddings from product descriptions.

The Cora dataset consists of scientific publications represented as nodes, with edges indicating citation links [51]. Node features are bag-of-words representations of document content, and the task is to assign each publication to its research topic.

The Patents dataset is a citation network of U.S. utility patents, where nodes represent patents and edges indicate citations [24, 26, 28]. Node features are derived from patent metadata, and the classification task is to predict the patent grant date.

A.2 Temporal road-scene graph construction

We treat road scene understanding as *risk-aware structured perception*: given a short temporal window, the model must infer semantic categories for traffic entities while quantifying uncertainty so that downstream modules, such as tracking, behaviour prediction, and planning, can handle ambiguous or novel situations conservatively rather than relying on overconfident predictions. Each temporal window is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$, where nodes $\mathcal{V}$ represent scene context and traffic entities, edges $\mathcal{E}$ encode spatial and temporal relations, and X denotes node features.

ROAD. The ROAD benchmark contains urban and suburban driving scenes with substantial variability in illumination, weather, traffic density, and interaction patterns [41]. Its object-level annotations and geometric context make it suitable for temporal scene-graph construction. ROAD is particularly relevant for uncertainty evaluation because it contains rare traffic agents and interaction patterns, such as occlusions, unusual merges, and complex intersection behaviour, which can naturally induce epistemic uncertainty. ROAD provides dense temporal annotations with 10 agent categories, 19 action attributes, and 12 location attributes. Each frame contributes one unlabeled scene node and one node for each annotated agent. Intra-frame edges are scene-centered: the scene node is connected bidirectionally to all agent nodes. Temporal edges connect scene nodes across adjacent frames and connect agent nodes with the same `tube_uid` across adjacent frames. Self-loops are also added. We use window size $= 4$, window stride $= 4$, frame step $= 5$. Node features concatenate bounding-box geometry, action one-hot features, location one-hot features, and node-type indicators. Agent-ID features are excluded to prevent leakage. The exported ROAD graphs use 37-dimensional node features.

nuScenes. We also evaluate on nuScenes, a large-scale autonomous-driving benchmark with 1000 scenes and synchronized multi-modal sensing [5]. The dataset provides six surround-view cameras, one LiDAR sensor, five radars, ego-pose and localization metadata, and map priors for supported regions. For nuScenes, we use 8 object categories: $\{$car, truck, bus, trailer, construction_vehicle, pedestrian, motorcycle, bicycle$\}$. Temporal windows use window size $= 3$, window stride $= 1$. The exported nuScenes graphs use 10-dimensional node features. We use the same temporal scene-graph formulation as for ROAD, while adapting feature construction and temporal-window parameters to the available annotations and sensing structure of nuScenes. Graph statistics for both temporal benchmarks are reported in Tab. 5. Fig. 4 illustrates, on a road-scene graph, how the softmax and belief heads respond to an agent of a type never seen in training.

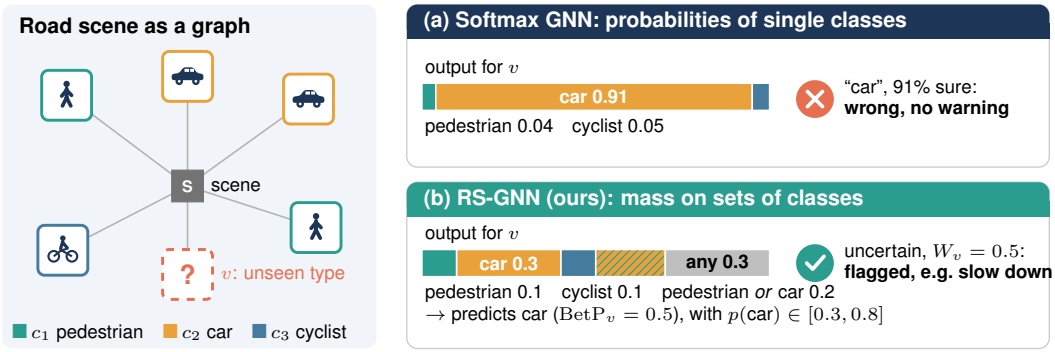


Figure 4: **Softmax vs. random-set head on a road-scene graph.** Agents are nodes connected through the scene node; the known classes are c_1 (pedestrian), c_2 (car), and c_3 (cyclist), and node v is an agent of a type never seen in training. *(a)* The softmax head must split all probability among the known classes and answers "car" with probability 0.91: a confident error with no warning. *(b)* RS-GNN assigns mass also to sets of classes (hatched: $\{c_1, c_2\}$; grey: any class). It still predicts c_2 ($\mathrm{BetP}_v = 0.5$), but only knows that $p(c_2) \in [0.3, 0.8]$: the width $W_v = 0.5$ flags the prediction as uncertain. Numbers are those of the worked example in Sec. 3.

Table 5: Temporal road-scene graph statistics used in the reported runs.

Dataset	#Train graphs	#Val graphs	#Test graphs	#Test labeled nodes
ROAD	49,123	26,302	26,301	291,197
nuScenes	307	39	38	6,677

B Cross-dataset behaviour and configuration sensitivity

Figure 5 provides a single-figure overview of all reported OOD AUROC values. On the primary score shown here, RS-GNN is the strongest method on Chameleon (84.18), ArXiv (70.21), and Roman Empire (69.58), and close to the best on Cora (88.84 vs. GNNSafe's 88.85). On Reddit2 and Coauthor, dedicated calibration baselines (CaGCN, GNNSafe) retain the lead. Per-dataset top-5 method comparisons are provided in Fig. 6; the rank of RS-GNN among all 14 methods is summarised in Fig. 7; and the in-distribution metrics on temporal road-scene benchmarks are visualised in Fig. 8.

C Extended ablation analysis

We ablate two design choices on **Cora** (small class space, full focal-set enumeration feasible) and **Roman Empire** (higher class count, requiring budgeted focal-set construction): focal-set construction (singletons-only vs. standard focal sets including non-singleton subsets) and loss formulation (full belief-space objective vs. binary cross-entropy with or without margin regularisation). The loss variants are: the full objective of Eq. (4); BCE only (no mass-validity terms); and BCE with margin regularisation.

The singletons-only configuration achieves the highest mean AUROC alongside relatively strong average accuracy. This suggests that even a restricted focal set can provide useful uncertainty estimates. However, the corresponding distribution exhibits a higher degree of variability across runs, indicating that this configuration is less stable. In this setting, the model is limited to assigning belief to individual classes, which restricts its ability to explicitly represent ambiguity between classes.

By contrast, the standard focal-set configuration as shown in Fig. 10 achieves the highest observed AUROC, while maintaining a more consistent distribution of results. Although its mean AUROC is slightly lower than that of the singletons-only configuration, the improved stability suggests that the additional expressiveness of the focal set enables more reliable modelling of uncertainty. In particular, the inclusion of multi-class subsets allows the model to capture partial knowledge in cases where evidence does not strongly support a single class.

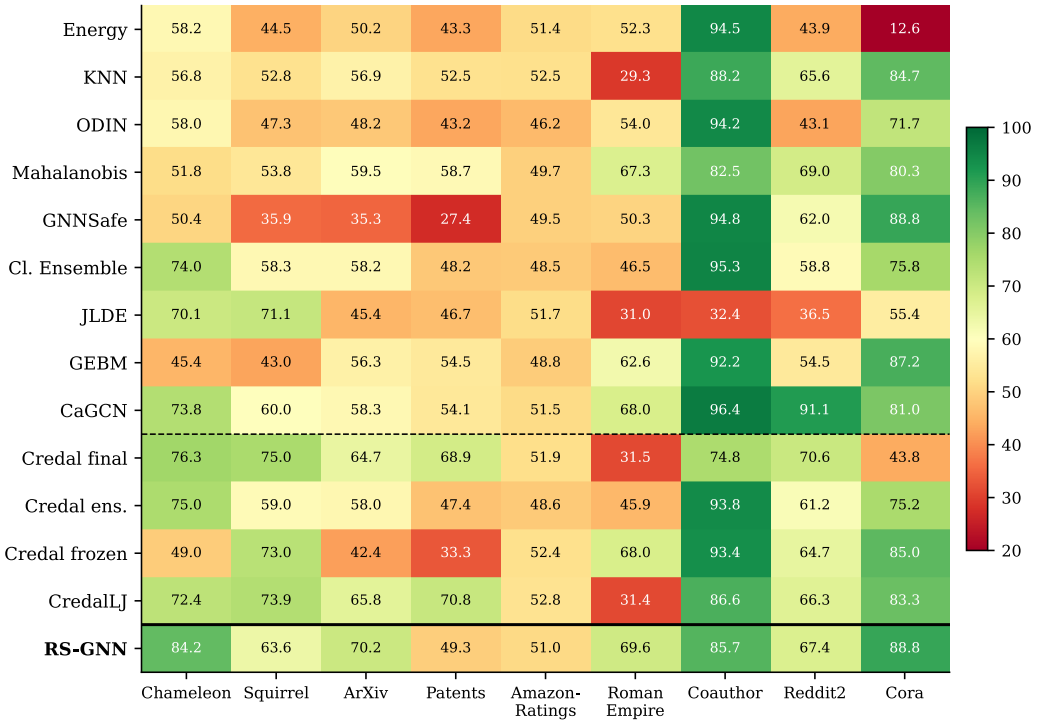


Figure 5: OOD detection AUROC (%) across all 14 methods and 9 standard graph benchmarks. The dashed line separates the post-hoc, graph-specific, and ensemble baselines (top) from the credal methods (middle); the solid line separates baselines from **RS-GNN** (bottom row, in bold). Cell colour encodes AUROC.

The BCE-based variants (standard focal sets with BCE and BCE with margin regularisation) demonstrate lower performance overall, with mean AUROC values of 0.6070 and 0.5777, respectively. These results suggest that the choice of loss function has a significant influence on the quality of the learned belief assignments. In particular, optimisation using standard probabilistic loss functions appears less effective in this context, as such losses are not explicitly designed to operate over set-valued representations. In contrast, the full training objective, which directly operates in belief space, is better aligned with the underlying uncertainty framework.

The results in Fig. 10 highlight the role of focal-set construction in balancing expressiveness and computational efficiency. While configurations using the full focal-set space achieve the highest peak AUROC, the performance gap relative to reduced or singleton-based focal sets remains relatively small. This suggests that the primary gains of RS-GNN can be attributed to the combinatorial richness of the focal-set family and to the underlying belief-based representation itself. In practical terms, this indicates that lightweight focal-set configurations may provide a more favourable trade-off between computational cost and uncertainty quality, particularly for large-scale datasets where full enumeration is infeasible. Another notable pattern is the divergence between classification accuracy and AUROC across configurations: although test accuracy remains consistently high, AUROC exhibits greater variability, reinforcing the need to evaluate uncertainty-aware models using metrics that explicitly capture their ability to distinguish between in-distribution and out-of-distribution samples.

The effect of the focal-set budget K on construction time and the inference-time cost of RS-GNN are reported in §D.8 (Tabs. 7 and 9).

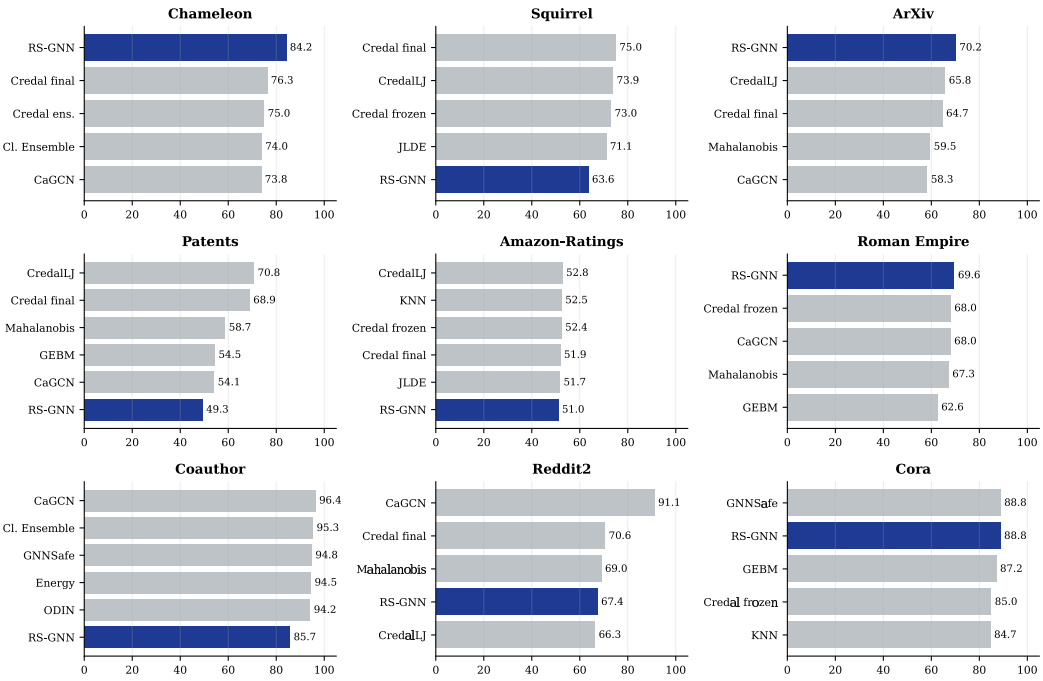


Figure 6: Per-dataset top-5 methods plus RS-GNN (highlighted in dark blue), by OOD detection AUROC (%). RS-GNN ranks first on Chameleon, ArXiv, and Roman Empire; second on Cora (0.01 behind GNNSafe); in the top-5 on Squirrel and Reddit2; and outside the top-5 on Patents, Amazon-Ratings, and Coauthor. The pattern aligns with the calibration-vs-representation distinction discussed in §5.

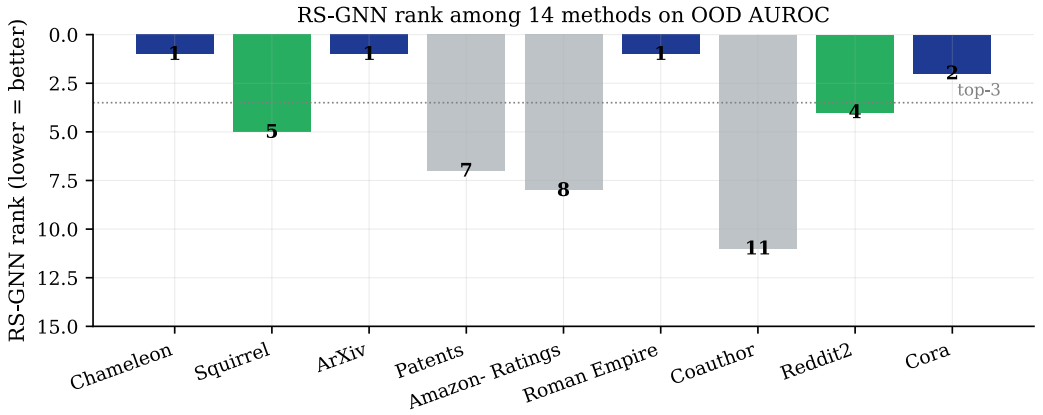


Figure 7: RS-GNN's rank (lower is better) among the 14 methods evaluated on OOD detection AUROC. RS-GNN reaches the top-3 on four of nine datasets (**Chameleon**, **ArXiv**, **Roman Empire**, and **Cora**) and is in the top-5 on Reddit2 and Squirrel; it ranks lower on Patents, Amazon-Ratings, and Coauthor. Its average rank is 4.44, second only to CaGCN (4.33).

D Implementation details

D.1 Models

We compare two models that share the same message-passing backbone and differ only in their output representation.

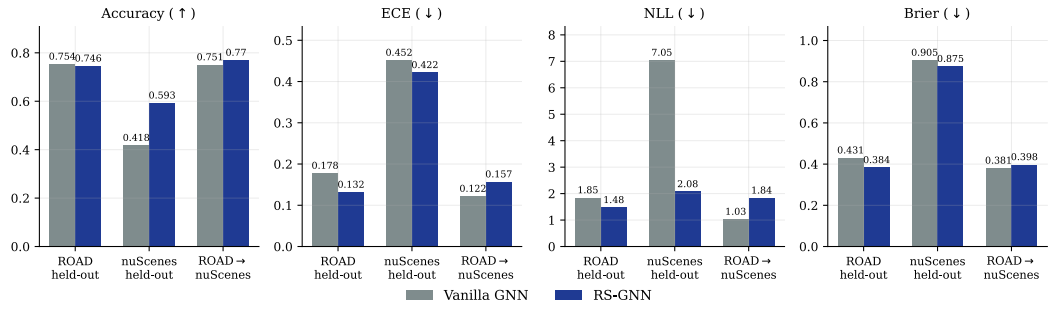


Figure 8: In-distribution metrics on the temporal road-scene benchmarks: Vanilla GNN (grey) vs. RS-GNN (blue). On nuScenes, RS-GNN improves accuracy from 0.418 to 0.593 and reduces NLL from 7.047 to 2.079; on the cross-dataset ROAD→nuScenes setting, vanilla retains a small ECE/NLL edge on ROAD ID metrics, reflecting the closed-set vs. uncertainty-aware trade-off.

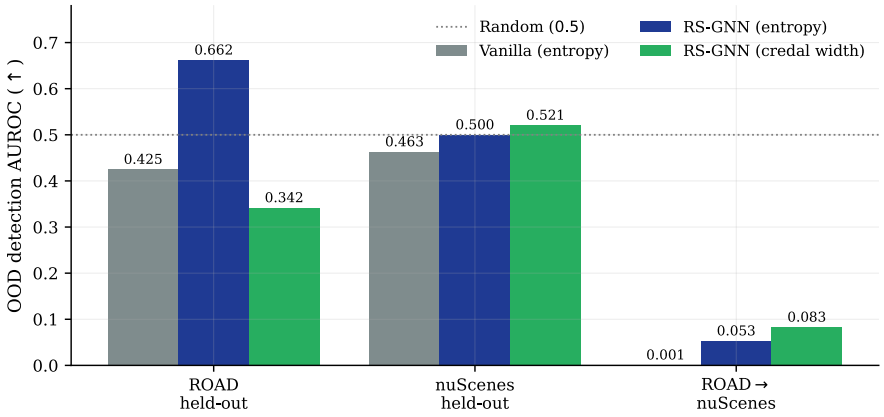


Figure 9: OOD detection AUROC across the three road-scene settings. Vanilla softmax entropy (grey) is weak on within-dataset shift and fully inverted on cross-dataset shift (AUROC ≈ 0.001). RS-GNN entropy (blue) improves on ROAD, where credal width (green) is inverted; both RS-GNN scores are near chance on nuScenes and less inverted than softmax under cross-dataset shift, where both models remain far below chance. The dotted line marks chance (AUROC 0.5).

Vanilla GNN (softmax head). A message-passing backbone computes node embeddings z_v, followed by a linear classifier and softmax, $p_v(i) = \mathrm{softmax}(Wz_v)_i$. This produces a point probability vector, and uncertainty is approximated using scalar confidence heuristics such as maximum softmax probability or predictive entropy. In closed-set settings this is effective; however, under strict leave-out-class shift, the softmax head must still assign probability mass to known singleton classes, which can lead to overconfident errors on unfamiliar OOD nodes.

RS-GNN (belief head). RS-GNN uses the same backbone but replaces the softmax head with a belief head over a focal-set budget $\mathcal{F}$. For each node v, the model predicts beliefs over focal sets using sigmoid activations. These beliefs are transformed into masses through Möbius inversion, and a pignistic projection yields a point prediction for accuracy and calibration evaluation. The key difference is representational: RS-GNN can allocate mass to non-singleton focal sets, explicitly encoding epistemic ambiguity when graph evidence is incomplete or conflicting.

D.2 Backbone and implementation details

Environment. Experiments use PyTorch, PyTorch Geometric, and PyTorch Lightning, run in the NVIDIA PyG container (`nvcr.io/nvidia/pyg:24.05-py3`, Dockerfile in the repository) or an equivalent Conda environment, with experiment tracking in Weights & Biases. All experiments were run on a server with 8 NVIDIA A100-SXM4 GPUs (40 GB each), two AMD EPYC 7742 CPUs (128 cores in total), and 1 TB of RAM, running Ubuntu 22.04 with CUDA 12.8; the experiments reported

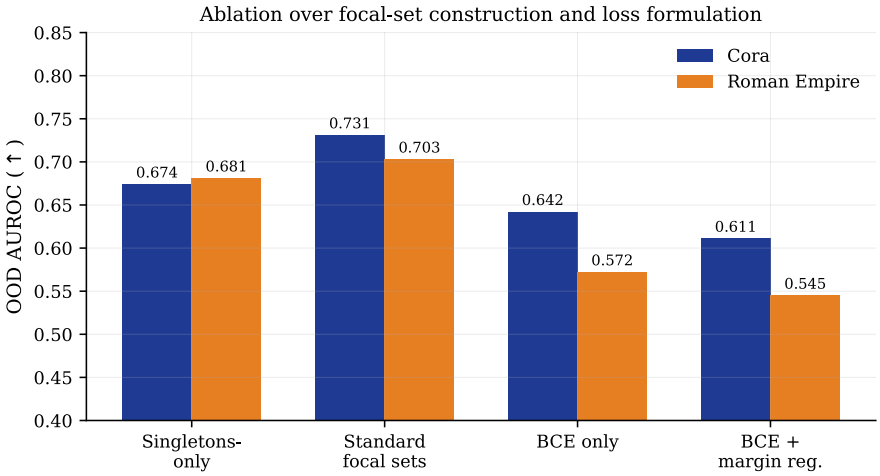


Figure 10: Ablation over focal-set construction and loss formulation, evaluated on Cora and Roman Empire. The standard focal-set configuration (singletons plus selected non-singleton subsets, full belief-space loss) attains the highest AUROC on both datasets; BCE-only and BCE+margin variants drop substantially (mean AUROC 0.607 and 0.578 respectively), indicating that probabilistic losses are not aligned with set-valued supervision.

in the paper required approximately 1,000 GPU hours in total. Experiments were executed via the command line:

Command: `python main.py -dataset <dataset_name> -model <model_name> -count <seeds>`

All models use `GATv2` [3] as the default message-passing backbone. GATv2 allows each node to aggregate a weighted combination of its neighbours' features, where attention coefficients are learned as a function of interacting node representations. This is particularly useful in road-scene graphs, where neighbourhoods may contain occlusions, spurious edges, short tracks, or rapidly changing context.

Concretely, the encoder is a two-layer `GATv2` network with hidden size 128. The first layer uses 4 attention heads, while the second layer collapses to a single head and returns a 128-dimensional embedding per node. ReLU is applied after each layer with dropout rate 0.2 after the first layer: $H^{(1)} = \sigma(\mathrm{GATv2}^{(1)}(X, \mathcal{E}))$, $H^{(2)} = \sigma(\mathrm{GATv2}^{(2)}(H^{(1)}, \mathcal{E}))$. The resulting node embedding is $z_v = H_v^{(2)} \in \mathbb{R}^{128}$.

Training is performed using NeighborLoader-based mini-batching, so each forward pass observes only a subgraph of the full dataset. This is compatible with RS-GNN because the random-set head does not require global graph normalisation or repeated stochastic forward passes. Each mini-batch produces a self-contained belief assignment over class subsets, allowing uncertainty estimation to remain local to the sampled neighbourhood.

D.3 Prediction heads

Both models share the same backbone and differ only in the prediction head.

In the vanilla model, logits are computed via a linear layer and converted to probabilities with softmax: $\ell_v = W z_v$, $p_v(i) = \mathrm{softmax}(\ell_v)_i$. In RS-GNN, the head outputs focal-set beliefs via a linear layer with sigmoid outputs, $\hat{\mathrm{Bel}}_v(A_k) = \mathrm{sigmoid}(g(z_v))_k$ for $A_k \in \mathcal{F}$. Masses are recovered by a precomputed Möbius inversion matrix M, $\hat{m}_v = \hat{\mathrm{Bel}}_v M$. Pignistic probabilities are then computed using the pignistic matrix P induced by $\mathcal{F}$, $\mathrm{BetP}_v = \mathrm{norm}(\hat{m}_v P)$, where $\mathrm{norm}(\cdot)$ denotes row-wise normalisation. Final predictions use $\arg\max_i \mathrm{BetP}_v(i)$, meaning RS-GNN differs from vanilla primarily in its uncertainty representation rather than its final decision rule.

RS-GNN belief head for node v

Input: node embedding $z_v \in \mathbb{R}^d$; focal family $\mathcal{F} = \{A_1, \ldots, A_M\}$; Möbius matrix M; pignistic matrix P
Output: pignistic probabilities BetP_v and credal-width score W_v

1. $a_v \leftarrow W z_v + b$
2. $\mathrm{Bel}_v \leftarrow \mathrm{sigmoid}(a_v)$ — *predicted beliefs over focal sets*
3. $m_v \leftarrow \mathrm{Bel}_v M$ — *Möbius inversion: beliefs → masses*
4. $m_v \leftarrow \max(m_v, 0)$ — *remove negative estimated masses*
5. $r_v \leftarrow \max\left(1 - \sum_{A \in \mathcal{F}} m_v(A), 0\right)$ — *unassigned residual mass*
6. If universal set $\mathcal{Y} \in \mathcal{F}$:
   (a) $m_v(\mathcal{Y}) \leftarrow m_v(\mathcal{Y}) + r_v$ — *assign residual to total ignorance*
7. If $\sum_{A \in \mathcal{F}} m_v(A) \leq \varepsilon$:
   (a) $\mathrm{BetP}_v \leftarrow (1/|\mathcal{Y}|, \ldots, 1/|\mathcal{Y}|)$ — *fallback for a degenerate output*
8. Else:
   (a) $m_v \leftarrow m_v / \sum_{A \in \mathcal{F}} m_v(A)$ — *normalize valid mass vector*
   (b) $\mathrm{BetP}_v \leftarrow m_v P$ — $\mathrm{BetP}_v(c) = \sum_{A \ni c} m_v(A)/|A|$
9. For each class $c \in \mathcal{Y}$:
   (a) $\mathrm{Bel}_v(c) \leftarrow \sum_{A \in \mathcal{F}: A \subseteq \{c\}} m_v(A)$
   (b) $\mathrm{Pl}_v(c) \leftarrow \sum_{A \in \mathcal{F}: c \in A} m_v(A)$
   (c) $w_v(c) \leftarrow \mathrm{Pl}_v(c) - \mathrm{Bel}_v(c)$
10. $W_v \leftarrow w_v(\hat{y}_v)$, with $\hat{y}_v = \arg\max_c \mathrm{BetP}_v(c)$ — *Eq. (3)*
11. Return BetP_v, W_v

Figure 11: RS-GNN belief head for node v.

The belief head $g_\theta : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{F}|}$ produces unconstrained logits followed by a coordinate-wise sigmoid, yielding independent belief scores in $(0,1)^{|\mathcal{F}|}$. Using sigmoid rather than softmax is dictated by our random-set learning principle: for a node of true class y, the target is multi-hot, $\mathrm{Bel}_{\mathrm{target}}(A) = \mathbb{1}(y \in A)$, meaning every focal set containing y is simultaneously consistent with the observation. Treating random-set prediction as a multi-label task allows mutually compatible propositions to co-exist without forcing them to compete in a softmax layer. Because exact mass validity is encouraged via soft regularizers $\mathcal{R}_{\mathrm{neg}}$ and $\mathcal{R}_{\mathrm{sum}}$ in the objective of Eq. (4) rather than strictly enforced, minor violations can occur during training and are penalized directly [31].

The budgeted Möbius inversion of Eq. (1) is the exact Möbius inverse when $\mathcal{F}$ is the full power set ($\mathcal{F} = 2^{\mathcal{Y}} \setminus \{\emptyset\}$).[1]

D.4 Hard-validity alternatives

Validity is currently encouraged through soft regularization during training and enforced at inference by clipping and renormalisation (Fig. 11). When the focal family is the full power set, the budgeted construction coincides with the exact random-set belief function. A hard-valid alternative during training is to project the raw mass vector onto the set of valid mass functions on $\mathcal{F}$ (a probability simplex over the selected focal sets),

$$\mathcal{M} = \Big\{ m \in \mathbb{R}^{|\mathcal{F}|} : \; m(A) \geq 0 \;\; \forall A \in \mathcal{F}, \;\; \sum_{A \in \mathcal{F}} m(A) = 1 \Big\}.$$

[1] On a budgeted focal family $\mathcal{F}$, the transformation is computed over the retained focal-set lattice and should be interpreted as a restricted approximation over those specific sets rather than a full reconstruction over all $2^{\mathcal{Y}}$ subsets. This formulation follows the shared property of the random-set model family [31].

Given unconstrained head outputs $\tilde{m}_v$, one can compute

$$m_v = \Pi_{\mathcal{M}}(\tilde{m}_v) = \arg\min_{m \in \mathcal{M}} \|m - \tilde{m}_v\|_2^2.$$

This projection enforces non-negativity and normalization for every node, so Proposition 1 would hold throughout training as well. Because $\mathcal{M}$ is a simplex, the projection can be computed in $O(|\mathcal{F}| \log |\mathcal{F}|)$ time and implemented as a differentiable simplex-projection layer. We currently use soft regularization to preserve training flexibility.

D.5 Proof of Proposition 1

For $B \subseteq \mathcal{Y}$ let $\mathrm{Bel}(B) = \sum_{A \in \mathcal{F}, A \subseteq B} m(A)$ and $\mathrm{Pl}(B) = \sum_{A \in \mathcal{F}, A \cap B \neq \emptyset} m(A)$. *(ii) and non-emptiness in (i).* $\mathrm{BetP}(c) = \sum_{A \ni c} m(A)/|A| \geq 0$ and $\sum_c \mathrm{BetP}(c) = \sum_A m(A) = 1$, so BetP is a distribution. For any B, $\mathrm{BetP}(B) = \sum_{A \in \mathcal{F}} m(A)\,|A \cap B|/|A| \geq \sum_{A \subseteq B} m(A) = \mathrm{Bel}(B)$, since every term is non-negative and equals $m(A)$ when $A \subseteq B$. Hence $\mathrm{BetP} \in \mathcal{P}(m)$ and $\mathcal{P}(m) \neq \emptyset$. *Convexity and closedness in (i).* $\mathcal{P}(m)$ is the intersection of the probability simplex with the finitely many closed half-spaces $\{p : p(B) \geq \mathrm{Bel}(B)\}$, each linear in p. *(iii).* Since $\emptyset \notin \mathcal{F}$, the only focal set contained in $\{\hat{y}\}$ is $\{\hat{y}\}$ itself, so $\mathrm{Bel}(\{\hat{y}\}) = m(\{\hat{y}\})$, while $\mathrm{Pl}(\{\hat{y}\}) = \sum_{A \ni \hat{y}} m(A)$. Their difference is $\sum_{A \ni \hat{y}, |A| > 1} m(A)$, a sum of non-negative masses bounded by $\sum_A m(A) = 1$. *(iv).* By (iii), $\mathrm{Pl}(\{c\}) - \mathrm{Bel}(\{c\}) = 0$ for all c if and only if $m(A) = 0$ for every A with $|A| > 1$, i.e. m is supported on singletons. In that case $\mathrm{Bel}(\{c\}) = m(\{c\})$ and, for $p \in \mathcal{P}(m)$, $p(c) \geq m(\{c\})$ for every c with $\sum_c p(c) = 1 = \sum_c m(\{c\})$, which forces $p(c) = m(\{c\}) = \mathrm{BetP}(c)$; thus $\mathcal{P}(m) = \{\mathrm{BetP}\}$. □

D.6 Focal-set construction

To keep random-set outputs scalable, we do not enumerate all of $2^{\mathcal{Y}}$. Instead, $\mathcal{F}$ contains all singleton sets over the full class space and up to 64 additional non-singleton sets of maximum cardinality 3, selected by the class-overlap procedure below. This yields 74 focal sets for ROAD and 72 for nuScenes (Tab. 6).

Table 6: Focal-set budgets used for RS-GNN. 'K non-singletons' is the number of non-singleton focal sets added to singleton focal sets.

Dataset	#Classes (total)	K non-singletons	#Total focal sets
ROAD	10 (ID=8, OOD=2)	64 (budgeted pairs/triples)	74
nuScenes	8 (ID=6, OOD=2)	64 (budgeted pairs/triples)	72

Focal-set selection is performed once as a preprocessing step, summarised in Alg. 1. For the standard benchmarks with fewer than 10 ID classes (all except Coauthor, Reddit2, and Roman Empire), $\mathcal{F}$ is the full non-empty power set of the ID classes (7 focal sets for 3 ID classes); the road-scene runs always use the budgeted construction (Tab. 6). Otherwise, up to 200 labelled ID training nodes per class are sampled, and their induced subgraph is embedded by a warm-up GNN with the backbone architecture, trained on labelled ID nodes only. The embeddings are projected to three dimensions with t-SNE, a single Gaussian is fitted to the points of each class, and each class is represented by its 95% confidence ellipsoid (χ^2_3 quantile 7.815), discretised on a regular 3-D grid. Candidate pairs, and triples extending pairs with non-zero overlap, are scored by the Jaccard overlap (intersection over union) of the discretised regions of their classes, and the K highest-scoring subsets with non-zero overlap are retained alongside all singletons. OOD labels are not used at any stage. Because only pairs and triples are considered, candidate generation has complexity $O(C^2 + C^3)$, followed by top-K selection; RS-GNN training and inference then use $C + K$ output heads.

For the focal-set budget ablation, the embedding, GMM-fitting, and ellipsoid-construction stages were executed once per dataset, after which the overlap-selection stage was evaluated three times for each budget K; no separate timing warm-up iterations were used. We used each dataset loader's default split (Roman Empire uses fixed split index 0). Candidate non-singleton focal sets were scored by the Jaccard overlap of their discretized ellipsoid regions, $|A \cap B|/|A \cup B|$, with a score of zero when the union was empty. Candidates were restricted to cardinality at most three, and the

K highest-scoring nonzero-overlap sets were retained alongside all singleton sets. Ties followed the underlying `torch.argsort` ordering. PyTorch and GMM fitting used seed 7; however, because t-SNE and all auxiliary random-number generators were not explicitly seeded, the procedure was not fully end-to-end deterministic.

Algorithm 1 Graph-aware focal-set construction (one-off preprocessing)

Require: graph $G = (V, E, X)$, labelled ID training nodes, ID classes $\mathcal{Y}$, budget K, maximum cardinality 3

if standard benchmark with $|\mathcal{Y}| < 10$ **then return** $\mathcal{F} = 2^{\mathcal{Y}} \setminus \{\emptyset\}$
end if
Sample up to 200 labelled ID training nodes per class; let G_s be their induced subgraph
$Z \leftarrow$ node embeddings of G_s from a warm-up GNN trained on labelled ID nodes; $Z_3 \leftarrow \mathrm{tSNE}_3(Z)$
for each class $c \in \mathcal{Y}$ **do**
    Fit a Gaussian $\mathcal{N}(\mu_c, \Sigma_c)$ to $\{Z_3[v] : y_v = c\}$; $R_c \leftarrow$ voxels inside its 95% ellipsoid
end for
$\mathcal{S} \leftarrow \emptyset$
for each pair $A = \{c, c'\}$, then each triple A extending a pair with non-zero score **do**
    $s(A) \leftarrow \left|\bigcap_{c\in A} R_c\right| / \left|\bigcup_{c\in A} R_c\right|$; add $(A, s(A))$ to $\mathcal{S}$
end for
return $\mathcal{F} = \{\{c\} : c \in \mathcal{Y}\} \cup$ top-K subsets of $\mathcal{S}$ with $s(A) > 0$

Algorithm 2 RS-GNN training and inference

Require: graph G, ID training labels, focal family $\mathcal{F}$ (Alg. 1), weights α, β

Precompute the Möbius matrix M and the pignistic matrix P of $\mathcal{F}$
for each epoch and mini-batch of labelled ID nodes B **do**
    $z_v \leftarrow \mathrm{GNN}_\phi(G)_v$; $\hat{\mathrm{Bel}}_v \leftarrow \sigma(g_\theta(z_v))$ for $v \in B$ ▷ message passing, belief head
    $\hat{m}_v \leftarrow \hat{\mathrm{Bel}}_v M$ ▷ Möbius inversion, Eq. (1)
    $\mathcal{L} \leftarrow \mathrm{BCE}(\mathrm{Bel}^{\mathrm{target}}, \hat{\mathrm{Bel}}) + \alpha\mathcal{R}_{\mathrm{neg}} + \beta\mathcal{R}_{\mathrm{sum}}$; update ϕ, θ ▷ Eq. (4)
end for
Inference: for each test node, compute $\hat{m}_v$, clip and renormalise it, and return $\mathrm{BetP}_v = \hat{m}_v P$, $\hat{y}_v = \arg\max_i \mathrm{BetP}_v(i)$, the pignistic entropy H_v, and the credal width W_v (Fig. 11)

D.7 Training objectives and hyperparameters

Vanilla training minimises cross-entropy on labeled ID nodes. RS-GNN minimises binary cross-entropy in belief space with mass-validity regularisation as defined in Eq. (4). For a labeled node with class y_v, the focal-set target is $\mathrm{Bel}_v^{\mathrm{target}}(A_k) = \mathbf{1}\{y_v \in A_k\}$. In the implementation, $\mathcal{L} = \mathrm{BCE}(\mathrm{Bel}^{\mathrm{target}}, \hat{\mathrm{Bel}}) + \alpha\,\mathbb{E}_{v,A}[\mathrm{ReLU}(-\hat{m}_v(A))] + \beta\,\mathbb{E}_v\big|\sum_{A_k\in\mathcal{F}} \hat{m}_v(A_k) - 1\big|$, where the first penalty discourages negative mass values and the second discourages deviations from unit total mass. In ROAD temporal graphs, unlabeled scene nodes with label -1 are excluded using $\mathcal{V}_{\mathrm{lab}} = \{v \in \mathcal{V} : y_v \geq 0\}$, and for strict leave-out-class OOD training the ID mask is $\mathcal{V}_{\mathrm{ID}} = \{v \in \mathcal{V}_{\mathrm{lab}} : y_v \notin \mathcal{Y}_{\mathrm{OOD}}\}$. Core hyperparameters for both runs are listed in Tab. 8.

D.8 Computational complexity

RS-GNN avoids the exponential 2^C output space by using C singleton classes and only K selected non-singleton focal sets, so the belief head costs $O(d(C + K))$ per node instead of $O(dC)$ for a softmax head, and the Möbius and pignistic transformations are fixed matrix products of size $|\mathcal{F}| \times |\mathcal{F}|$ and $|\mathcal{F}| \times C$. The one-time focal-set construction considers candidate pairs and triples, with complexity $O(C^2 + C^3)$, followed by top-K selection. We report two complementary measurements. First, for $K \in \{4, 8, 16, 32, 64\}$, focal-set construction time remains essentially unchanged (Tab. 7), confirming that the dominant cost is the fixed embedding and overlap computation rather than the budget size; increasing K therefore mainly changes the output dimension $C+K$. This cost is incurred once, before training, and never at inference. Second, Tab. 9 reports the additional forward-pass time

Table 7: Focal-set construction time as a function of the focal-set budget K. The number of focal sets includes the C singleton sets. Construction is performed once during training and is not incurred at inference.

Dataset	K	# Focal Sets	Estimated Total Time (s)
Coauthor	4	15	21.679878 ± 0.029825
	8	19	21.651761 ± 0.002800
	16	27	21.655925 ± 0.013529
	32	43	21.649985 ± 0.003718
	64	75	21.652124 ± 0.004639
Reddit2	4	34	42.756976 ± 0.024109
	8	38	42.674700 ± 0.079591
	16	46	42.704531 ± 0.023639
	32	62	42.704572 ± 0.060480
	64	94	42.987422 ± 0.243270
Roman-Empire	4	17	23.237838 ± 0.055103
	8	21	23.246564 ± 0.007441
	16	29	23.251319 ± 0.017514
	32	44	23.294706 ± 0.057734
	64	77	23.315260 ± 0.027014

Table 8: Core hyperparameters used in the two hold-out-class runs.

Hyperparameter	ROAD	nuScenes
Backbone	GATv2 (2 layers)	GATv2 (2 layers)
Hidden dim	128	128
Attention heads	4, 1	4, 1
Dropout	0.2	0.2
Batch size	16	16
Optimizer	Adam	Adam
Learning rate	1×10^{-3}	5×10^{-4}
Epochs	30	80
α, β	$10^{-3}, 10^{-3}$	$10^{-3}, 10^{-3}$

of RS-GNN and of post-hoc OOD detectors over the same backbone: RS-GNN adds 0.07–2.7 ms, with a lower mean overhead than Energy and Mahalanobis on every dataset and roughly 80 to 6,000 times less than ODIN and KNN.

Table 9: Additional per-forward-pass runtime of RS-GNN and post-hoc OOD baselines relative to the same frozen GNN backbone. Values are mean $\pm$ sample standard deviation in milliseconds across the evaluated checkpoints.

Dataset	RS-GNN additional time (ms)	Energy additional time (ms)	ODIN additional time (ms)	KNN additional time (ms)	Mahalanobis additional time (ms)
Amazon-Ratings	0.370 ± 0.049	1.391 ± 0.508	60.359 ± 3.867	229.383 ± 31.893	1.615 ± 1.376
ArXiv	2.705 ± 0.279	10.924 ± 9.625	440.586 ± 21.547	$16{,}332.308 \pm 224.526$	10.101 ± 5.859
Chameleon	0.069 ± 0.008	0.563 ± 0.689	28.661 ± 0.956	6.183 ± 0.898	0.824 ± 1.190
Coauthor	0.487 ± 0.007	6.260 ± 6.986	714.208 ± 354.840	266.344 ± 16.922	7.487 ± 0.685
Cora	0.080 ± 0.021	0.290 ± 0.281	24.072 ± 8.167	8.183 ± 1.506	1.153 ± 0.206
Roman-Empire	0.651 ± 0.069	1.312 ± 0.274	50.036 ± 1.254	225.063 ± 9.886	11.852 ± 0.257
Squirrel	0.110 ± 0.025	0.675 ± 4.596	90.539 ± 0.742	17.400 ± 0.318	0.572 ± 1.256

D.9 Transductive experimental setup

Nodes belonging to the held-out OOD classes remain in the training graph, but their labels are masked and they are excluded from the supervised objective. Consequently, their features and connectivity can participate in message passing, while no OOD-class supervision is provided. This is the standard transductive setting used in our comparison and is applied identically to RS-GNN and

all baselines, with the same graph, data split, and backbone. Therefore, any benefit or degradation caused by unlabeled OOD nodes propagating information through the graph is shared across methods. Transductive exposure may make OOD detection easier when held-out nodes form structurally distinct clusters, but it may also make it harder when their messages contaminate ID representations, particularly in heterophilic graphs.

Table 10: Dataset statistics and ID/OOD partitions. The table reports graph size, number of classes, graph homophily regime, class assignments, and the number of ID/OOD nodes in the training, validation, and test splits. Ho and He denote homophilic and heterophilic datasets, respectively.

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
Cora	2,708	5,429	7	Ho	{0,1,2,3}	{4,5,6}	421	167 / 333	316 / 684
Amazon-Ratings	24,492	93,050	5	He	{0,1}	{2,3,4}	4,461	2,231 / 3,892	2,230 / 3,893
Roman-Empire	22,662	32,927	18	He	{0,…,4}	{5,…,17}	5,216	2,609 / 3,056	2,608 / 3,058

D.10 Feature sparsity across evaluation benchmarks

We interpret sparse node features as node representations in which only a small fraction of feature dimensions are nonzero. We computed this statistic for all existing benchmarks and found that our evaluation already covers a wide range of feature regimes. In particular, Chameleon, Squirrel, Coauthor, Cora, and Patents are highly sparse, with feature sparsity ranging from 98.1% to 99.4%. Chameleon has only 0.55% nonzero entries and 10.2% of its nodes have an all-zero feature vector, while Patents contains exactly 5 nonzero features per node out of 269 dimensions. Conversely, Reddit2, Roman-Empire, ArXiv, and Amazon-Ratings have nearly dense features. Comparing with Tab. 1, RS-GNN performs well in both regimes: among sparse datasets it ranks first on Chameleon and second on Cora but lower on Coauthor and Patents, and among dense datasets it ranks first on ArXiv and Roman Empire but lower on Amazon-Ratings. We therefore do not attribute performance differences to feature sparsity, since topology, homophily, scale, and class count also vary across datasets; a controlled feature-masking study is left to future work.

Table 11: Node-feature sparsity statistics for the evaluated datasets. Mean NNZ/node denotes the average number of nonzero features per node; density and sparsity report the percentage of nonzero and zero entries, respectively; zero-feature nodes denotes the proportion of nodes with no active input features.

Dataset	# Features	Mean NNZ/node	Density (%)	Sparsity (%)	Zero-feature nodes (%)
Chameleon	2,325	12.81	0.551	99.449	10.23
Squirrel	2,089	17.97	0.860	99.140	3.17
Coauthor	6,805	59.57	0.875	99.125	0.00
Cora	1,433	18.17	1.268	98.732	0.00
Patents	269	5.00	1.859	98.141	0.00
Reddit2	602	599.80	99.635	0.365	0.00
Roman-Empire	300	299.18	99.726	0.274	0.27
ArXiv	128	128.00	99.9998	0.0002	0.00
Amazon-Ratings	300	300.00	100.000	0.000	0.00

D.11 Evaluation metrics

We evaluate both models on labeled test nodes. For vanilla, probabilities are softmax outputs p_v; for RS-GNN, probabilities are pignistic probabilities $\mathrm{Bet}\mathrm{P}_v$, and the predictive uncertainty is the pignistic entropy of Eq. (2). The vanilla baseline uses analogous softmax entropy and maximum-softmax-probability scores, $s_v^{\mathrm{MSP}} = 1 - \max_i p_v(i)$. For binary OOD detection, where $t_v = \mathbf{1}\{y_v \in \mathcal{Y}_{\mathrm{OOD}}\}$, we report AUROC, AUPRC, and FPR95. For RS-GNN we additionally report the predicted-class credal width W_v of Eq. (3). Calibration is evaluated using ECE (15 bins), NLL, and Brier score: $\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} |\mathrm{acc}(S_b) - \mathrm{conf}(S_b)|$, $\mathrm{NLL} = -\frac{1}{N}\sum_{v=1}^{N} \log(p_v(y_v) + \varepsilon)$, and $\mathrm{Brier} = \frac{1}{N}\sum_{v=1}^{N}\sum_{i=0}^{C-1}(p_v(i) - \mathbf{1}\{y_v = i\})^2$.

Table 12: Test ID ECE and OOD AUROC (pignistic entropy) of RS-GNN before and after post-hoc isotonic calibration. Results are reported as mean $\pm$ sample standard deviation over five random seeds.

Dataset	ECE Before	ECE After	Δ ECE	AUROC Before	AUROC After	Δ AUROC
Amazon-Ratings	0.1511 ± 0.0877	0.0063 ± 0.0033	-0.1449	0.4836 ± 0.0020	0.4914 ± 0.0095	$+0.0077$
ArXiv	0.0261 ± 0.0203	0.0088 ± 0.0041	-0.0174	0.5227 ± 0.0347	0.5345 ± 0.0090	$+0.0118$
Chameleon	0.0703 ± 0.1014	0.0497 ± 0.0297	-0.0206	0.5602 ± 0.1354	0.5284 ± 0.1305	-0.0317
Coauthor	0.0984 ± 0.0431	0.1118 ± 0.0597	$+0.0134$	0.5149 ± 0.0929	0.5057 ± 0.0804	-0.0092
Cora	0.1422 ± 0.0953	0.0543 ± 0.0289	-0.0879	0.6274 ± 0.1496	0.6117 ± 0.1199	-0.0157
Roman-Empire	0.0513 ± 0.0454	0.0136 ± 0.0081	-0.0377	0.4498 ± 0.1319	0.3853 ± 0.0692	-0.0645
Squirrel	0.0156 ± 0.0104	0.0270 ± 0.0089	$+0.0113$	0.5197 ± 0.0281	0.5433 ± 0.0263	$+0.0236$
ROAD	0.0464 ± 0.0174	0.0346 ± 0.0376	-0.0119	0.5527 ± 0.0761	0.5262 ± 0.0641	-0.0265
nuScenes	0.2117 ± 0.0630	0.1001 ± 0.0157	-0.1116	0.4464 ± 0.0710	0.4990 ± 0.0247	$+0.0525$

Post-hoc isotonic calibration is applied to the pignistic probabilities produced by a trained RS-GNN. Isotonic regressors are fitted on validation nodes and applied at test time. Results are reported as mean $\pm$ sample standard deviation over five random seeds. Calibration improves test ID ECE on seven of nine datasets, although its effect on entropy-based OOD detection is dataset-dependent. For Coauthor, ECE increases slightly from 0.0984 ± 0.0431 to 0.1118 ± 0.0597, while OOD AUROC decreases from 0.5149 ± 0.0929 to 0.5057 ± 0.0804. In contrast, on nuScenes, ECE improves from 0.2117 ± 0.0630 to 0.1001 ± 0.0157, accompanied by an AUROC increase from 0.4464 ± 0.0710 to 0.4990 ± 0.0247. On ROAD, ECE similarly improves from 0.0464 ± 0.0174 to 0.0346 ± 0.0376, whereas AUROC decreases from 0.5527 ± 0.0761 to 0.5262 ± 0.0641. Overall, AUROC improves on Amazon-Ratings, ArXiv, Squirrel, and nuScenes, and decreases on Chameleon, Coauthor, Cora, Roman-Empire, and ROAD. Thus, isotonic calibration is a lightweight complementary method that generally improves ID probability calibration, but does not uniformly improve OOD detection.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state the method (belief head, graph-aware focal sets, credal-width uncertainty) and the empirical findings; these are supported by Tabs. 1–3 and qualified in Sec. 5, including where RS-GNN is outperformed and the fact that cross-dataset shift is not solved.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: Sec. 5 discusses calibration-bottlenecked datasets, the fixed focal-set selection and its budget, extreme cross-dataset shift, configuration sensitivity, the transductive protocol, and soft mass validity during training; computational cost and scaling are analysed in Sec. 3 and §D.8.

   Guidelines:

   - The answer [N/A] 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: Proposition 1 states its assumptions (a valid mass function on a focal family not containing the empty set), and its complete proof is given in §D.5.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The method is fully specified in Sec. 3, with pseudocode in Algs. 1–2 and Fig. 11; datasets, ID/OOD partitions, and splits are in Tab. 10 and §A; architecture, hyperparameters, and focal-set budgets are in §D (Tabs. 6 and 8).

   Guidelines:

   - The answer [N/A] 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: All datasets are publicly available, and the code, with a README, the exact commands (§D.2), and the sweep configurations, is released at the repository linked in Sec. 1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The protocol, baselines, and backbone are described in Sec. 4; data splits are in Tab. 10; architecture, optimizer, and hyperparameters are in §D (Tab. 8); the hyperparameter sweep configurations are included in the released code.

   Guidelines:

   - The answer [N/A] 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: [No]

Justification: The main OOD-detection and road-scene results (Tabs. 1–3) are reported as single values without error bars; run-to-run variability is discussed in Sec. 5, and error bars (sample standard deviation over five seeds or repeated measurements) are reported for the isotonic-calibration study and the cost measurements (Tabs. 12, 7, and 9).

Guidelines:

- The answer [N/A] 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: §D.2 describes the hardware (8 NVIDIA A100-SXM4 40 GB GPUs, two AMD EPYC 7742 CPUs, 1 TB RAM) and reports that the experiments required approximately 1,000 GPU hours in total; per-component costs are in Tabs. 7 and 9.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research uses only publicly available benchmark datasets, involves no human subjects, and conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: Positive impacts (safer, uncertainty-aware graph predictions in safety-critical settings) and negative impacts (over-reliance on uncertainty estimates that can miss some distribution shifts) are discussed in the Broader impact paragraph of Sec. 6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with a high risk of misuse.

    Guidelines:

    - The answer [N/A] 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: All datasets, baselines, and software libraries are credited with citations (Sec. 4, §A, §D); they are publicly available for research use and are used in accordance with their terms of use (e.g., the non-commercial research licences of ROAD and nuScenes).

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The released code is documented with a README describing installation, the experiment commands, and the configuration options.

Guidelines:

- The answer [N/A] 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: [N/A]

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

Guidelines:

- The answer [N/A] 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: [N/A]

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

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not a component of the proposed method or of the experiments; they were used only for editing the manuscript.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.