DOES NODE-LEVEL OUT-OF-DISTRIBUTION DETECTION NEED STOCHASTIC UNCERTAINTY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Node-level out-of-distribution (OOD) detection is usually addressed with stochastic uncertainty methods, but the benchmarks that rank them have received little scrutiny. We show that the most widely reused protocols reward input scale and graph-level contrast rather than uncertainty. Standard Gaussian feature noise inflates feature norms on row-normalised graphs, and scoring a synthetic graph against the original tests graph discrimination instead of node detection. We propose a controlled protocol suite with scale-matched node-level corruptions, neighbourhood rewiring, held-out classes, and a real temporal shift, together with DCE, a deterministic encoder whose classifier reads each node's normalised features beside its aggregated context. Once scale is matched, sampling barely changes detection, whereas DCE outperforms dedicated uncertainty methods on feature anomalies in a single forward pass and achieves excellent accuracy. We also prove that node-wise decoders under separable losses cannot exploit a structured prior's cross-node correlation, and our experiments show no gain from such priors.

1 INTRODUCTION

A node classifier on an evolving graph must also indicate which of its predictions are unreliable, and the graph literature addresses this need with dedicated uncertainty methods. MC-Dropout and deep ensembles sample weights or train several networks (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), evidential models propagate Dirichlet evidence (Zhao et al., 2020; Stadler et al., 2021), GNNSafe propagates the energy of a deterministic classifier (Wu et al., 2023a), latent graph SDEs evolve node states (Bergna et al., 2025), and structured latent priors make the injected noise follow the topology (Rue & Held, 2005; Yoo et al., 2022). These methods are compared through node-level out-of-distribution (OOD) detection, in which a score derived from the classifier must rank anomalous test nodes above the rest. The anomalies are nodes whose features are overwritten by noise, nodes of classes withheld in training, or the nodes of a synthetic graph generated from the original (Stadler et al., 2021; Wu et al., 2023a).

However, the two most widely reused designs allow a detector to succeed without estimating uncertainty. GPN overwrites node features with standard Gaussian noise (Stadler et al., 2021), and, when the clean features are row-normalised, this noise far exceeds their norm (Figure 1a). Consequently, a classifier that feeds raw features into its first layer becomes more confident on the corrupted nodes, as ReLU networks do far from their data (Hein et al., 2019). Its entropy ranks them below clean ones, and a propagated energy score such as GNNSafe can fall far below chance. Disagreement between sampled predictions responds oppositely, and here the mutual information of MC-Dropout over a graph attention network outperforms every other score we evaluate. GNNSafe's synthetic shifts avoid the scale problem, but they score a generated graph against the original in a second pass (Wu et al., 2023a), which rewards the discrimination of two graphs rather than the detection of anomalous nodes within one. Each effect vanishes under a targeted change that leaves the remainder of the protocol intact. A linear projection with LayerNorm (Ba et al., 2016) can raise GNNSafe from near zero to well above chance. Noise drawn from the empirical feature moments reduces the mutual information of MC-Dropout from the best score to chance, and confining this noise to a subset of the test nodes in a single graph degrades propagated energy more than any other score.

We redesign the benchmark along the two axes that these designs conflate, namely the scope of the shift and the corruption that produces it, and we add neighbourhood rewiring, held-out classes,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and a real temporal shift. Once the scale channel is closed, sampling no longer improves detection, because averaging over dropout masks or ensemble members smooths a predictive distribution without changing which nodes the underlying network finds ambiguous. In contrast, DCE, a deterministic context encoder whose read-out sees each node's own normalised representation beside the aggregated context, outperforms the best dedicated score by a wide margin on every scale-matched feature protocol in one forward pass, and it achieves excellent accuracy against a broad set of baselines. Its advantage vanishes when an anomaly leaves a node's own features intact, since rewiring favours two-layer backbones and held-out classes favour propagated energy over entropy. No score reliably detects the temporal shift, whose newer nodes are only marginally harder to classify.

We also show that a standard classifier cannot exploit the cross-node correlation of a structured latent prior. When the logits at a node read only that node's latent draw and the loss is a sum over nodes, priors with equal row marginals induce identical per-node predictive laws, risks, and expected gradients, and a rank-k spectral prior stays within average total-variation distance $O(k/n)$ of isotropic noise through every measurable decoder of that kind. GCNF attaches a graph-conditional normalising flow with a swappable spectral prior to DCE and tests the claim where this certificate is no longer small. Relative to its isotropic control, the structured prior changes no accuracy, detection, or risk-coverage endpoint at full rank, under a pairwise objective, behind a latent-mixing decoder, or with all three together, and removing the noise barely affects detection. Since a null result cannot distinguish an unused prior from a model never trained to use one, we restrict these claims to the pattern the theorem describes.

Our contributions are threefold. We show that the two most reused node-level OOD designs reward input scale and graph-level contrast and turn them into a controlled suite. We introduce DCE, a deterministic encoder that detects node-level feature anomalies better than dedicated uncertainty methods and achieves excellent accuracy, trace that margin to a read-out of each node's own normalised features, and identify the unsolved shifts. We prove that a node-wise decoder under a separable objective cannot see a structured prior's correlation shape and sees a low-rank prior as nearly isotropic noise, and with GCNF we show that the prior stays inert even when each premise is removed.

2 Related Work

Uncertainty Estimation for Node Classification. Message passing (Kipf & Welling, 2017; Velickovic et al., 2018; Hamilton et al., 2017), decoupled propagation (Gasteiger et al., 2019), graph transformers (Wu et al., 2022; 2023b; Shirzad et al., 2023), continuous diffusion (Chamberlain et al., 2021), and masked pretraining (Li et al., 2023) emit one logit vector per node, and uncertainty is added on top. Model-agnostic remedies sample dropout masks or train ensembles (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017), Bayesian treatments sample the graph or its edge masks (Zhang et al., 2019; Hasanzadeh et al., 2020), evidential models fit Dirichlet posteriors (Zhao et al., 2020; Stadler et al., 2021; Damke & Hüllermeier, 2024), energy detectors score logits (Liu et al., 2020) and propagate or regularise that energy over the graph (Wu et al., 2023a; Yang et al., 2024), and latent SDEs evolve node states (Bergna et al., 2025). Calibration (Guo et al., 2017; Wang et al., 2021; Hsu et al., 2022) and conformal prediction (Zargarbashi et al., 2023; Huang et al., 2023) assess confidences and prediction sets instead.

Evaluation of OOD Detection. GPN evaluates on leave-out classes and on per-node feature perturbations with Gaussian or Bernoulli noise (Stadler et al., 2021), a design that later evidential work extends (Damke & Hüllermeier, 2024). GNNSafe instead scores an OOD graph generated by feature interpolation or stochastic-block-model edges against the original, alongside leave-out classes and real shifts on Twitch and ogbn-arxiv (Wu et al., 2023a), and later detectors adopt its splits (Yang et al., 2024; Wang et al., 2025). ReLU networks are known to grow overconfident far from the data (Hein et al., 2019). To our knowledge, no prior work has measured how much of the published ranking of node-level detectors this overconfidence and the graph-level contrast of synthetic OOD graphs explain, and we measure both in Section 3.

Structured Latent Models and Normalising Flows. Structured variational autoencoders bring GMRF structure (Rue & Held, 2005; Yoo et al., 2022) to the per-node Gaussians of Kipf & Welling (2016), and flows reach graphs through molecular generation (Madhawa et al., 2019; Shi et al., 2020) and reversible message passing (Liu et al., 2019). Of twenty-two models sorted in Appendix D, only

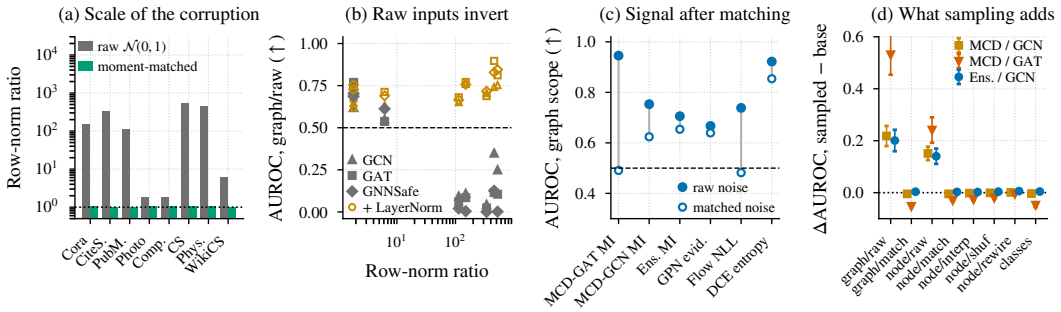

Figure 1: Scale effects of Gaussian feature noise, as ten-trial means. (a) Row-norm ratio, corrupted over clean. (b) Raw-input detectors against it. (c) Raw against matched noise. (d) Gain from sampling.

GCNF, which we build as an instrument, decodes node-wise and trains separably, and our negative result on priors therefore applies to that pattern alone.

3 What the Reused Protocols Measure

3.1 Node-Level Detection

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with n nodes, adjacency matrix A, degree matrix $D = \mathrm{diag}(A\mathbf{1})$, features $X \in \mathbb{R}^{n \times f}$, and labels $y \in \{1, \dots, C\}^n$ observed on $\mathcal{V}_{\mathrm{tr}}$. A classifier returns logits $\hat{y}_v \in \mathbb{R}^C$ and a categorical distribution $p_v = \mathrm{softmax}(\hat{y}_v)$ at each node, and a detector ranks test nodes by a score s_v read off them. The score can be the predictive entropy $\mathcal{H}(p_v) = -\sum_c p_{vc} \log p_{vc}$, the energy $E_v = -\log \sum_c \exp \hat{y}_{vc}$ of the logits (Liu et al., 2020) or its propagated form in GNNSafe, the mutual information between prediction and sampled weights or latents (Houlsby et al., 2011; Gal et al., 2017), or GPN's total evidence. A protocol alters a set $\mathcal{O}$ of test nodes, and detection is the AUROC of s_v with $\mathcal{O}$ as the positive class, where values below 0.5 mark an inverted score. The *scope* of a protocol is the graph when every test node is altered and a corrupted pass is scored against a clean one, and it is the node when a fifth of the test nodes are altered and ranked against the rest in one forward pass. Its *corruption* is the change that makes the nodes of $\mathcal{O}$ anomalous.

3.2 The Scale Channel

Raw corruption draws $\boldsymbol{x}_v \sim \mathcal{N}(\mathbf{0}, I)$, GPN's Gaussian perturbation, and its effect depends on the scaling of the real features. A row-normalised bag-of-words vector spreads unit mass over its few active words, so its Euclidean norm is small, whereas a standard Gaussian row has norm close to $\sqrt{f}$. On our five citation and co-authorship graphs the corrupted rows are therefore two to three orders of magnitude longer than the clean ones (Figure 1a), whereas the binary counts of the Amazon graphs and the embeddings of WikiCS have norms within an order of magnitude of $\sqrt{f}$.

A network that feeds raw features into a linear layer carries this scale into its logits. The logits of a network built from linear maps and ReLU-type activations grow roughly linearly along a ray from the origin, so a much longer input generically yields a nearly one-hot softmax, the regime in which ReLU networks become arbitrarily confident (Hein et al., 2019). Entropy, and energy whenever some logit grows along the ray, then rate the corrupted nodes as more familiar than the clean ones, and at graph scope plain GCN, GAT, and GNNSafe invert on exactly the five row-normalised graphs but not on the Amazon graphs (Figure 1b). The failure arises before any message passing and is independent of the graph. A linear projection with LayerNorm removes it at the same stage and raises GNNSafe from near zero to well above chance. LayerNorm divides each projected row by its own spread, so the next layer sees the shape of the features rather than their length. With the projection in place, raw and moment-matched noise give similar detection, so scale accounts for most of what raw noise adds.

Scores that measure spread respond in the opposite direction. On an inflated input each dropout mask removes large activations, so the logits swing between passes and MC-Dropout's mutual information reads the input scale as disagreement. At graph scope it becomes the best of the twenty-three scores we evaluate, and the latent likelihood of the flow of Section 6.3 also responds strongly. Matched noise draws feature j from $\mathcal{N}(\hat{\mu}_j, \hat{\sigma}_j^2)$ with the moments of the training nodes, which keeps the row norm

Table 1: Detection AUROC (↑) by entropy, or GNNSafe's propagated energy, as the mean over eight benchmarks of ten-trial means ± the s.d. across them. **Red bold** is best, blue underline second.

Method	graph/raw	graph/matched	node/raw	node/matched	node/interp	node/shuffle	node/rewire	classes
GCN	0.329 ± 0.226	0.672 ± 0.046	0.319 ± 0.179	0.598 ± 0.054	0.592 ± 0.051	0.587 ± 0.048	0.819 ± 0.070	0.825 ± 0.073
GAT	0.307 ± 0.305	0.691 ± 0.041	0.322 ± 0.187	0.593 ± 0.052	0.590 ± 0.048	0.587 ± 0.044	0.860 ± 0.072	0.846 ± 0.068
GCN (LN)	0.689 ± 0.054	0.670 ± 0.044	0.604 ± 0.049	0.596 ± 0.050	0.596 ± 0.048	0.594 ± 0.046	0.838 ± 0.080	0.814 ± 0.067
GAT (LN)	0.743 ± 0.073	0.690 ± 0.042	0.617 ± 0.052	0.594 ± 0.049	0.596 ± 0.047	0.597 ± 0.046	**0.871 ± 0.072**	0.841 ± 0.075
MC-Dropout (GCN)	0.547 ± 0.150	0.667 ± 0.049	0.470 ± 0.109	0.593 ± 0.056	0.590 ± 0.051	0.586 ± 0.047	0.820 ± 0.066	0.822 ± 0.073
MC-Dropout (GAT)	0.838 ± 0.112	0.639 ± 0.041	0.563 ± 0.089	0.563 ± 0.046	0.564 ± 0.040	0.567 ± 0.040	0.855 ± 0.071	0.798 ± 0.079
Deep Ensemble	0.530 ± 0.173	0.675 ± 0.048	0.459 ± 0.128	0.601 ± 0.056	0.595 ± 0.052	0.591 ± 0.050	0.825 ± 0.072	0.829 ± 0.072
GNNSafe	0.269 ± 0.312	0.702 ± 0.037	0.277 ± 0.184	0.561 ± 0.043	0.555 ± 0.041	0.550 ± 0.039	0.791 ± 0.048	**0.860 ± 0.085**
GNNSafe (LN)	0.748 ± 0.059	0.713 ± 0.029	0.568 ± 0.044	0.564 ± 0.038	0.561 ± 0.037	0.559 ± 0.036	0.811 ± 0.053	0.848 ± 0.081
GPN	0.549 ± 0.022	0.606 ± 0.051	0.539 ± 0.014	0.574 ± 0.033	0.574 ± 0.037	0.567 ± 0.042	0.728 ± 0.055	0.811 ± 0.068
GAT-deep (LN)	0.896 ± 0.048	0.754 ± 0.088	0.831 ± 0.079	0.716 ± 0.098	0.697 ± 0.087	0.707 ± 0.082	0.825 ± 0.091	0.799 ± 0.106
DCE (ours)	0.922 ± 0.049	**0.854 ± 0.073**	0.903 ± 0.050	0.838 ± 0.078	**0.786 ± 0.077**	**0.782 ± 0.072**	0.829 ± 0.089	0.813 ± 0.088
GCNF (DCE + flow)	**0.935 ± 0.042**	0.852 ± 0.079	**0.924 ± 0.040**	**0.841 ± 0.080**	0.781 ± 0.080	0.779 ± 0.073	0.816 ± 0.091	0.811 ± 0.085

close to its clean value but still destroys the co-occurrence structure of real features, and it reduces both scores to chance (Figure 1c). Both detect the scale of the input rather than uncertainty about the class. The same change reverses the apparent value of sampling, which raises GAT far above its deterministic counterpart under raw noise and lowers it under matched noise (Figure 1d).

3.3 A Controlled Protocol Suite

We retain the raw setting as a reference and close each confound separately (Appendix B). Four corruptions run at node scope. Raw and matched noise are as above. Interpolation follows GNNSafe in setting $\boldsymbol{x}_v = \lambda \boldsymbol{x}_a + (1-\lambda)\boldsymbol{x}_b$ for training nodes a and b with $\lambda \sim \mathrm{U}(0,1)$ (Wu et al., 2023a), so it stays inside the convex hull of real features. Shuffling permutes the feature vectors of $\mathcal{O}$ randomly among themselves and leaves almost no node in place, so only a node's mismatch with its neighbourhood can reveal it. Raw and matched noise also run at graph scope. Two further protocols leave features untouched. Structural rewiring gives each node in $\mathcal{O}$ as many new edges as it had, to uniformly random endpoints, so almost none of the original incident edges remain. The semantic protocol withholds a third of the classes from training and validation while their nodes stay in the graph (Stadler et al., 2021). The last arm is a real shift in which models train on the ogbn-arxiv papers published up to 2017 and score every later paper in the same pass (Hu et al., 2020; Wu et al., 2023a).

4 A Deterministic Baseline

If the raw protocols reward sampling only through the scale channel, a deterministic encoder under which an anomalous node still stands out after normalisation should match or outperform the sampled methods on the controlled suite. DCE, the deterministic context encoder, tests this hypothesis. It consists only of standard components and also conditions the flow of Section 6.3.

Network Architecture. The input map projects each node's features to width d_h, normalises them, and adds a gated positional encoding computed from the eigenvectors U_k of the k smallest strictly positive eigenvalues of the normalised Laplacian $\tilde{\boldsymbol{L}} = I - D^{-1/2}AD^{-1/2}$ (Belkin & Niyogi, 2003),

$$H^{(0)} = \mathrm{LN}(XW_h) + e^{\gamma}\rho(U_k), \tag{1}$$

with LayerNorm LN (Ba et al., 2016), a two-layer MLP ρ, and a learnable gate e^{γ}. Each of K layers then updates the representation residually through M-head graph attention (Velickovic et al., 2018),

$$H^{(\ell)} = \mathrm{LN}\big(H^{(\ell-1)} + \mathrm{GATConv}_M(H^{(\ell-1)}, \mathcal{G})\big), \qquad \ell = 1, \ldots, K, \tag{2}$$

and a two-layer MLP ψ reads the first and the last layer together, $p_v = \mathrm{softmax}\big(\psi([H_v^{(K)} \| H_v^{(0)}])\big)$, in the manner of jumping-knowledge read-outs (Xu et al., 2018). DCE trains on label-smoothed cross-entropy and scores node v by $\mathcal{H}(p_v)$ from one evaluation-mode pass. MC-Dropout needs thirty passes and our ensemble five networks, although each of their passes runs a smaller network.

Design Rationale. The normalised projection closes the scale channel before any aggregation, so the detector must see content rather than scale, while depth, width, and the recipe allow that content to separate from clean features. At node scope most neighbours of an anomalous node are clean, and the read-out of $H_v^{(0)}$ gives the head the node's own representation before aggregation mixes it with that clean context. The same reasoning predicts that the read-out cannot help under rewiring, because a rewired node keeps clean features and its direct path carries no sign of the corrupted neighbourhood.

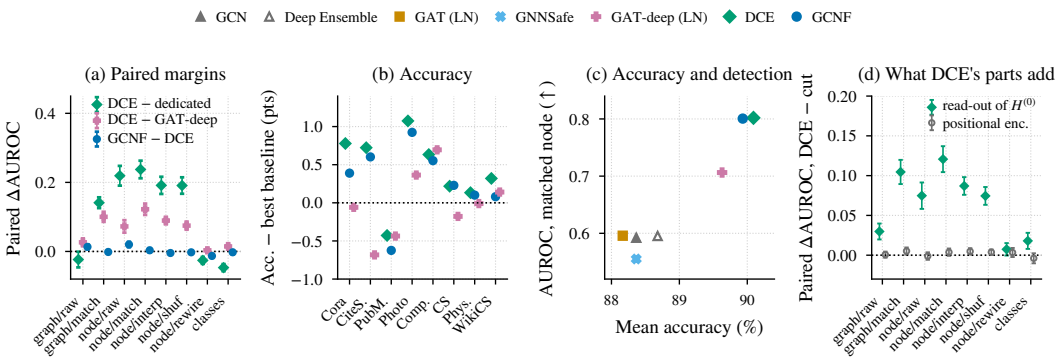

Figure 2: Main results over ten trials per benchmark. (a) Paired margins over 80 trial-graph cells with 95% intervals. (b) Accuracy over the best of eleven baselines. (c) Accuracy against matched node-level detection. (d) Cost of removing each distinguishing part of DCE, paired as in (a).

5 Detection Without Sampling

5.1 Experimental Setup

Datasets. We evaluate on eight benchmarks: the citation graphs Cora, CiteSeer, and PubMed (Sen et al., 2008; Namata et al., 2012), the co-purchase graphs Amazon Photo and Amazon Computers and the co-authorship graphs Coauthor CS and Coauthor Physics (Shchur et al., 2018), and the hyperlink graph WikiCS (Mernyei & Cangea, 2020). Each trial draws a class-stratified $60/20/20$ split (Luan et al., 2023; Liao et al., 2026). The temporal arm uses ogbn-arxiv (Hu et al., 2020), and Table 3 lists every graph.

Baselines. We compare accuracy against GCN, GraphSAGE, GAT, APPNP, GRAND, GPN, NodeFormer, DIFFormer, Exphormer, MaskGAE, and LGNSDE. The detection comparators are two-layer GCN and GAT scored by entropy, the same backbones behind a linear projection with LayerNorm, MC-Dropout over each with $S = 30$ passes, a five-member ensemble of GCNs, GNNSafe with and without the projection, and GPN. Since DCE is deeper and wider, GAT-deep (LN) adds a four-layer residual GAT that shares DCE's width and recipe and lacks only the positional encoding and the read-out of $H^{(0)}$. Each model is also read by every other score it supports (Appendix B).

Implementation and Statistical Testing. DCE uses $K = 4$ attention layers of width $d_h = 128$ and $k = 64$ eigenvectors, and it trains under the recipe of Appendix A.3, which it shares with the context track of GCNF without separate tuning. Each benchmark has ten trials that fix split and initialisation, and ogbn-arxiv has three. We report population standard deviations, test contrasts pairwise over the 80 trial-graph cells of a protocol with Holm correction (Holm, 1979), and read nulls as equivalence against prespecified margins where stated (Schuirmann, 1987; Lakens, 2017).

5.2 Main Results

OOD Detection Performance. Table 1 orders the protocols from the least to the most controlled. Once scale is matched, every two-layer method falls within one narrow band, and within it a dedicated score differs little from the entropy of its backbone. This is expected if these scores are functionals of a single predictive distribution, since averaging over dropout masks, averaging over ensemble members, and propagating energy all reshape a distribution whose ranking of nodes the backbone has already fixed. DCE rises above that band on every scale-matched protocol and outperforms every dedicated score on every feature protocol except raw graph-level noise, where the mutual information of MC-Dropout still benefits from the inflated input (Figure 2a). The gain lies in the encoder rather than the score, since DCE is scored by plain entropy from one deterministic pass. The ordering reverses once the shift leaves a node's own features intact, as rewiring favours the two-layer GAT (LN) and held-out classes favour propagated energy (Section 5.3). No score reliably detects the temporal shift of ogbn-arxiv, where every default score stays below 0.55 AUROC (Appendix B.1). Its newer papers are barely harder to classify, so scores of predictive difficulty find little signal.

Classification Accuracy. As reported in Table 2 and Figure 2b, DCE achieves excellent accuracy across the benchmarks, so its detection margin incurs no loss in accuracy. The uncertainty-aware

Table 2: Accuracy (%), mean±s.d. over ten trials. **Red bold** is best, blue underline second.

Method	Cora	CiteSeer	PubMed	Amazon Photo	Amazon Computers	CS	Physics	WikiCS
GCN	86.94 ± 1.53	74.83 ± 0.65	88.52 ± 0.52	93.88 ± 0.67	90.71 ± 0.50	92.97 ± 0.53	96.08 ± 0.25	82.83 ± 0.77
GraphSAGE	86.82 ± 1.29	74.26 ± 0.82	**89.91 ± 0.57**	94.94 ± 0.60	90.66 ± 0.77	94.06 ± 0.35	96.64 ± 0.17	85.19 ± 0.63
GAT	87.57 ± 1.13	75.24 ± 0.74	86.98 ± 0.64	93.48 ± 0.87	89.41 ± 1.02	92.79 ± 0.41	95.70 ± 0.28	82.39 ± 0.58
APPNP	87.76 ± 1.26	74.96 ± 1.20	86.33 ± 0.49	94.40 ± 0.44	89.73 ± 0.29	87.56 ± 0.28	94.92 ± 0.21	81.41 ± 0.84
GRAND	85.28 ± 1.53	74.23 ± 0.77	86.69 ± 0.43	94.83 ± 0.32	89.76 ± 0.37	95.47 ± 0.32	94.38 ± 0.32	81.21 ± 0.55
GPN	85.57 ± 1.73	73.93 ± 1.28	85.95 ± 0.69	89.22 ± 0.75	77.88 ± 2.20	87.79 ± 1.85	94.51 ± 0.61	76.88 ± 1.83
NodeFormer	75.44 ± 1.65	73.37 ± 1.25	88.39 ± 0.28	91.56 ± 0.65	86.20 ± 0.59	94.46 ± 0.34	96.27 ± 0.29	81.61 ± 0.74
DIFFormer	86.93 ± 1.09	75.48 ± 1.32	88.91 ± 0.45	93.81 ± 0.54	88.14 ± 0.30	94.11 ± 0.67	96.86 ± 0.22	80.44 ± 1.10
Exphormer	82.56 ± 1.26	68.61 ± 1.49	88.93 ± 0.56	94.90 ± 0.44	91.41 ± 0.42	89.65 ± 1.41	92.00 ± 1.76	83.99 ± 0.72
MaskGAE	85.50 ± 1.32	72.55 ± 0.75	86.51 ± 0.47	93.63 ± 0.60	89.85 ± 0.38	93.74 ± 0.21	95.96 ± 0.17	82.09 ± 0.58
LGNSDE	87.33 ± 1.29	75.66 ± 0.63	87.01 ± 0.42	94.04 ± 0.65	91.14 ± 0.25	93.86 ± 0.40	96.33 ± 0.30	82.92 ± 0.78
DCE (Ours)	**88.54 ± 1.35**	**76.39 ± 0.37**	89.48 ± 0.54	**96.01 ± 0.28**	**92.05 ± 0.38**	95.69 ± 0.28	**96.99 ± 0.19**	**85.51 ± 0.60**
GCNF (DCE + flow)	88.15 ± 1.27	76.27 ± 0.98	89.28 ± 0.55	95.86 ± 0.38	91.97 ± 0.29	**95.70 ± 0.33**	96.96 ± 0.16	85.27 ± 0.62

baselines do not convert their uncertainty estimates into higher accuracy. LGNSDE remains close to its underlying GCN, and GPN attains the lowest accuracy on both co-purchase graphs. The graph transformers perform competitively on some benchmarks but degrade on others, whereas DCE maintains its performance across all four graph domains under a single shared configuration. Moreover, DCE attains this accuracy in one deterministic forward pass without sampling, ensembling, or an evidential head. Detection and accuracy also improve jointly across methods (Figure 2c), consistent with one representation that separates both classes and atypical feature patterns.

5.3 What Each Part of DCE Contributes

Effect of Input Normalisation. The projection with LayerNorm removes the inversion, and label smoothing does not. Retraining plain GCN and GAT with DCE's label smoothing raises their entropy on clean nodes but leaves it near zero on inflated ones, so on the five row-normalised graphs their mean AUROC under raw graph-level noise falls from 0.169 to 0.128 for GCN and from 0.077 to 0.039 for GAT (Appendix B.2). Smoothing tempers confidence only near the training data, whereas the logits keep growing along the ray. The projection alone lifts both backbones above chance everywhere, yet it changes nothing under matched noise, so the repair is specific to scale (Table 1).

Effect of Depth, Width, and the Read-Out. GAT-deep (LN) separates capacity from design, and capacity and recipe account for much of DCE's lead on the feature protocols. The remainder survives correction on every one of them (Figure 2a) and is attributable to the read-out alone. Removing the read-out costs DCE almost exactly this remainder, whereas removing the positional encoding changes no protocol after correction (Figure 2d). The read-out places a node's own features beside its aggregated context. Shuffling only permutes real feature vectors within $\mathcal{O}$, so a score of each node's own features alone is at chance on average, and the read-out's gain there must come from comparing them with the context. Rewiring also sets features against context, but its random neighbours point to no single rival class, so deep residual encoders, which anchor each prediction to the node's own clean features, remain confident and detect rewiring less reliably than the two-layer GAT (LN).

Choice of Detection Score. Disagreement and density scores track the scale channel rather than the anomaly. On scale-matched node-level protocols every mutual-information, evidence, and density score trails its own model's entropy, since perturbing the weights or latents of an in-range input yields disagreement that carries less information than the mean prediction. Propagated energy instead measures the total evidence in the logits rather than its split among classes and smooths it over the graph, which favours held-out classes, whose nodes carry little evidence for a known class and cluster in homophilous regions. On DCE it raises held-out-class detection from 0.813 to 0.848, close to GNNSafe's 0.860 (Table 7). At graph scope the smoothing pools the shift of each altered neighbourhood without touching the clean nodes, which a separate pass scores, whereas at node scope both share one pass and the smoothing leaks each anomaly into its clean neighbours. Under matched noise, moving from graph to node scope therefore costs every backbone's propagated energy 0.09 to 0.15 AUROC, against less than 0.08 for its entropy.

Effect of Sampling. Once scale is matched, sampling leaves the entropy ranking of the underlying network essentially unchanged (Figure 1d), and LGNSDE, whose latent paths follow a graph-convolutional SDE, stays within 0.025 AUROC of the plain GCN that drives its dynamics under every scale-matched protocol on the three graphs where we score it (Appendix B). The small but consistent gain of the ensemble matches the variance reduction expected from averaging independently trained networks, at five times the cost, whereas MC-Dropout over GAT, whose attention dropout also perturbs the aggregation, degrades detection. A network trained with dropout scores one mask per

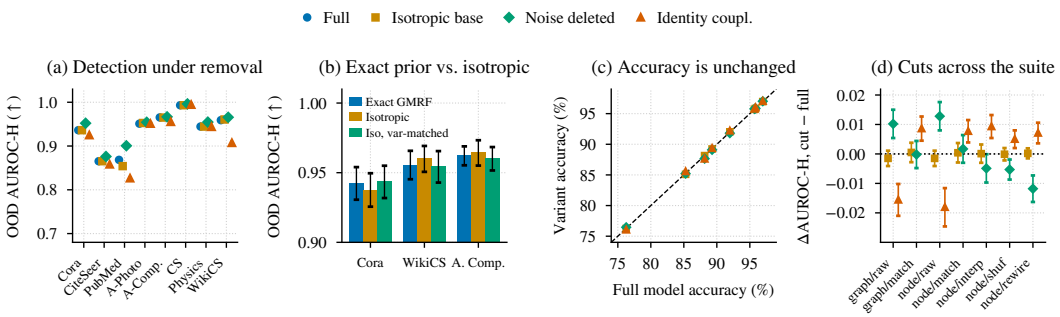

Figure 3: Cutting GCNF's stochastic path, with every variant retrained. (a) Detection under raw graph-level noise. (b) The full-rank prior against isotropic controls. (c) Each variant's accuracy against the full model's. (d) Each cut under the seven perturbation protocols, with 95% paired intervals.

step, so its loss penalises any variation of the true-class probability across masks (Appendix E), whereas ensemble members train separately and are never penalised for disagreement. Relative to DCE, the flow of GCNF, which also scores one draw per step, helps only on the two raw protocols and costs detection under rewiring (Figure 2a). Both effects grow once its draw is deleted (Section 6.4), so they arise from the flow's deterministic path rather than from sampling.

6 Can a Structured Latent Prior Help?

We prove that a node-wise decoder cannot see the cross-node correlation of a structured latent prior and test this result with GCNF.

6.1 Node-Wise Decoders and Spectral Priors

A stochastic node classifier draws a latent matrix $N \in \mathbb{R}^{n\times d}$ from a joint law P that may depend on the graph and maps it to logits. It is *node-wise* when the logits at v depend on row v of the draw alone,

$$\hat{y}_v = G_v(\boldsymbol{n}_v), \qquad \boldsymbol{n}_v := N_{v,:}, \tag{3}$$

for a measurable G_v fixed by the features, the graph, and the weights. The condition constrains the latent rather than the graph, since G_v may propagate or attend across the whole topology and only a draw that travels between nodes is excluded. Training is *separable* when the objective is a sum of per-node terms, $\mathcal{L}(\boldsymbol{\theta}; N) = |\mathcal{V}_{\mathrm{tr}}|^{-1} \sum_{v\in\mathcal{V}_{\mathrm{tr}}} \ell_{v,\boldsymbol{\theta}}(\boldsymbol{n}_v)$, as cross-entropy is. The canonical graph-aware prior is a Gaussian Markov random field, a matrix normal whose d channels share a row precision $I + \beta\tilde{\boldsymbol{L}}$ that favours smooth graph signals (Rue & Held, 2005),

$$p_0(Z \mid \mathcal{G}) = \mathcal{MN}_{n\times d}\Big(\mathbf{0},\ (I+\beta\tilde{\boldsymbol{L}})^{-1},\ I_d\Big), \quad \beta > 0. \tag{4}$$

Exact sampling costs $O(n^3)$ for a dense graph, and polynomial filters approximate the whole operator (Hammond et al., 2011), but a cheaper surrogate keeps only the k smallest strictly positive modes of $\tilde{\boldsymbol{L}} = U\Lambda U^\top$, which positional encodings already cache (Chung, 1997). We draw $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ and set

$$\boldsymbol{z}_{\mathrm{sp}} = \boldsymbol{\varepsilon} + U_k \operatorname{diag}(\boldsymbol{\delta})\, U_k^\top \boldsymbol{\varepsilon}, \qquad \delta_i = (1+\beta\lambda_i)^{-1/2} - 1, \tag{5}$$

which matches Eq. (4) on the retained modes and acts as the identity elsewhere, at cost $O(nkd)$ and with a closed-form log-determinant (Appendix E.1). We write $u_{vi} = [U_k]_{v,i}$ and $w_i = \beta\lambda_i/(1+\beta\lambda_i) \in (0,1)$ for the damping of mode i, call this law the low-rank spectral surrogate $\tilde{p}_0^{(k)}$, and reserve GMRF for the full-rank setting $k = n - c$ with c edge-containing components. Truncation keeps the smooth modes, which have the smallest w_i, so the surrogate at $k = 64$ damps little, leaves the rough modes untouched, and tilts the noise away from smooth signals, unlike the full field.

6.2 What a Node-Wise Decoder Can See

Eq. (3) makes a node-wise prediction read the prior through one row marginal, and Appendix C proves what follows for the objective and for unequal marginals.

Proposition 6.1 (Equal-marginal risk invariance and spectral marginals)**.** *Fix the parameters and any auxiliary randomness ξ the decoder carries, such as training dropout masks, drawn independently of the latent, and suppose that conditionally on ξ the decoder is node-wise in the sense of Eq.* (3) *and the graph-dependent part of G_v evolves independently of the draw. Let P and P' be two joint noise laws with equal row marginals, both held fixed with respect to the decoder parameters. Then they induce the same per-node predictive laws and the same expectation of every separable risk whose per-node terms are integrable. If the architectural hypothesis holds over an open parameter domain, then, wherever differentiation and expectation commute, their expected decoder gradients also agree, whereas their joint predictive laws and stochastic-gradient covariances need not. For the prior of Eq.* (5) *the row marginal is $\mathcal{N}(\mathbf{0}, \sigma_v^2 I_d)$, nondegenerate at every node, with*

$$\sigma_v^2 = 1 - \sum_{i=1}^{k} w_i\, u_{vi}^2, \qquad \left|\sigma_v^2 - 1\right| \le m_v := \sum_{i=1}^{k} u_{vi}^2, \qquad \frac{1}{n}\sum_v m_v = \frac{k}{n}.$$

The cross-node correlation *shape* of the prior therefore contributes nothing to per-node laws at fixed weights, and what remains splits into a scalar *budget* $B := 1 - n^{-1}\sum_{i\le k} w_i$ shared by all nodes and a node-wise *profile* $\pi_v = \sigma_v^2 - B$, with $1 - B \le k/n$ for any rank-k damping filter (Proposition C.3).

Theorem 6.2 (Decoder-independent rank bottleneck)**.** *Fix the graph, features, and decoder weights, and let $\mu_v^{(k)}$ and $\mu_v^{(0)}$ denote the output laws at node v under the rank-k prior of Eq.* (5) *and an isotropic $\mathcal{N}(\mathbf{0}, I_d)$ prior, respectively. Under the node-wise hypothesis of Eq.* (3)*, with σ_v^2 the row-marginal variance of Proposition 6.1, set*

$$a_v := 1 - \sigma_v^2 = \sum_{i=1}^{k} w_i u_{vi}^2, \qquad \rho := \max_i w_i < 1.$$

Write $\mathrm{TV}(P, Q) := \sup_A |P(A) - Q(A)|$. Then, for every measurable node-wise decoder, regardless of its sensitivity,

$$D_{\mathrm{KL}}\left(\mu_v^{(k)} \middle\| \mu_v^{(0)}\right) \le \frac{d}{2}\{-a_v - \log(1 - a_v)\}, \tag{6}$$

$$\frac{1}{n}\sum_{v=1}^{n} \mathrm{TV}\left(\mu_v^{(k)}, \mu_v^{(0)}\right) \le \sqrt{\frac{d}{8(1-\rho)}}\,(1 - B) \le \rho\sqrt{\frac{d}{8(1-\rho)}}\,\frac{k}{n}. \tag{7}$$

Consequently Eq. (7) *also bounds the average change in the expectation of any node-wise statistic with values in $[0, 1]$, and without an incoherence assumption on U_k the k/n rate cannot be improved.*

At $k = 64$, $d = 32$, and trained β, the certified distance from isotropic noise stays below 0.28% on every benchmark (Appendix F). Both results hold at fixed weights, so we retrain every intervention of Section 6.4. Low rank, node-wise decoding, and separability are assumptions, so we remove each individually and jointly, and Proposition C.4 extends the certificate to decoders that pool bounded sets of latent rows.

6.3 GCNF: A Stochastic Flow on Top of DCE

Testing the certificate in regimes where it is no longer small requires a prior that can be swapped, rescaled, or truncated while the decoder stays fixed, and GCNF provides one by attaching a conditional normalising flow with an exact latent likelihood to DCE (Appendix E). Its stochastic track starts from a second normalised projection plus the spectral noise,

$$\boldsymbol{z}^{(0)} = \phi_z(X) + \boldsymbol{z}_{\mathrm{sp}}, \qquad \phi_z(X) = \mathrm{LN}(XW_z) \in \mathbb{R}^{n\times d}, \tag{8}$$

which is the only place the prior enters. Each of K affine couplings (Dinh et al., 2017) rescales and shifts half of $\boldsymbol{z}_v^{(\ell-1)}$ by functions of the other half and of the DCE states $H_v^{(\ell)}$ and $H_v^{(0)}$ of the same node (Eq. (37)), and the head reads $\hat{y}_v = \psi([\boldsymbol{z}_v^{(K)} \| H_v^{(K)} \| H_v^{(0)}])$. The DCE states never read the draw, and the draw at v reaches only the prediction at v, so GCNF is node-wise. It trains on

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_z \mathcal{L}_{\mathrm{aux}}, \qquad \lambda_z = 1, \tag{9}$$

where $\mathcal{L}_{\text{aux}}$ is the same cross-entropy on a linear probe of $\boldsymbol{z}_v^{(K)}$, so the objective is separable and its expectation under any joint noise law factorises over the row marginals (Eq. (38)). Point predictions use one pass at the base mode $\boldsymbol{z}_\star^{(0)} = \phi_z(X)$, and uncertainty scores $S = 30$ draws by predictive entropy and by their mutual information with the prediction. It adds a $d = 32$ latent and four couplings to DCE and monitors its latent likelihood without optimising it (Appendices A.3 and E).

6.4 What Retraining Around the Prior Changes

Ablation of the Stochastic Path. Three retrained cuts isolate the role of each part of the stochastic path (Figure 3). We score detection under raw graph-level noise, where GCNF detects better than DCE, and panel (d) repeats the cuts under every perturbation protocol of the suite. Replacing the rank-64 base with isotropic noise changes nothing under any protocol, consistent with the certificate, and three graphs return identical accuracies on every trial. Under raw noise, identity couplings cost detection on some graphs, which could mean that the transport carries dependence that a node-wise head cannot otherwise reach, or only that it absorbs the perturbation the sampler injects. Removing the draw supports the latter reading, since detection then improves on every graph and decoding the trained full model at base mode does not reproduce the gain. Each training step scores a single draw, and among predictions with the same average over draws the expected log loss is lowest for one that does not vary between draws, by the strict concavity of the logarithm (cf. Bishop, 1995). Training therefore favours this absorption, since it penalises every use of a draw that carries no label information, whatever the prior's shape (Appendix E). Elsewhere in the suite every cut moves detection by less than 0.02, in a direction set by the protocol and independent of the prior (Figure 3d), and the flow adds no accuracy over DCE (Figure 2b) at 2.2 to 7.8 times DCE's parameters (Appendix H.1).

Full-Rank Prior. Retraining on every positive spectral mode of Cora, WikiCS, and Amazon Computers lowers the mean marginal variance to about a fifth, where the certificate no longer binds and the prior does alter what each node sees. It still changes no endpoint relative to an isotropic control, a variance-matched control that keeps its average scale but not its node profile, a unit-budget rescaling that keeps its shape and relative profile, or a neighbour-conditioned coupling that exposes its correlations (Figure 3b), and paired equivalence tests declare 23 of the 27 contrast cells equivalent after Holm correction (Appendix F.1). A last control draws independent rows at each node's exact marginal variance, so the two arms differ only in the cross-node correlation that Proposition 6.1 rules out at fixed weights. After retraining, no corrected difference test approaches significance, although only four of its nine cells are precise enough to certify equivalence (Appendix F.2).

Non-Separable Objective. Every arm above keeps a separable loss, so a null result might only mean that the objective never rewards the prior. A pairwise composite log score over training edges (Varin et al., 2011) supplies that incentive, because it rewards covariance, across draws, of the true-class probabilities at the two ends of an edge, and we cross it with the neighbour-conditioned decoder and the structured or variance-matched prior. The pairwise term raises the realised dependence, most strongly when both premises are crossed. Even the largest realised dependence stays more than thirty times below the 0.01-nat margin of the pairwise endpoint. The prespecified interaction fails on every graph, and 51 of the 60 contrast cells are equivalent after Holm correction (Appendix G.1), consistent with a latent channel too weak for its correlations to change any endpoint. On a block graph where cross-node dependence is the only learnable signal, averaging independent noise over neighbours recovers 25% of the pairwise likelihood that only such dependence can earn, whereas a node-wise read-out of the structured prior recovers 8% to 21% as the damping grows, since it can correlate neighbours only by damping the prior's rough modes. Together they recover less than the sum of their shares, so the routes are substitutes and the cheaper route needs no spectral prior (Appendix G.2).

7 Conclusion

In this paper, we show that once benchmarks stop rewarding input scale and graph-level contrast, node-level OOD detection needs a stronger encoder rather than stochastic uncertainty. DCE, which reads each node's normalised features beside its context, detects feature anomalies better than dedicated uncertainty methods in one deterministic pass and achieves excellent accuracy, while sampling and node-wise structured priors add almost nothing. Rewiring, held-out classes, and temporal shift remain open, and our nulls cannot separate an unused prior from one never trained to be used. Future benchmarks should match corruption moments and score anomalies inside a single graph.

AI USE STATEMENT

We used generative AI tools to assist with implementation of the method and evaluation pipeline, refinement of hypotheses and their framing, design of the experimental protocol, and polishing the presentation and formulation of mathematical claims. All mathematical claims and AI-assisted work were reviewed and verified by the authors. The authors take responsibility for the final content of the submission.

REPRODUCIBILITY

All datasets are public and are loaded through standard libraries. The appendices specify the data processing, the model configurations, the training procedures, the baseline settings, the software and hardware, and complete proofs of the theoretical claims. Everything needed to reproduce the reported results is in the paper and in the anonymized source code submitted as supplementary material.

ETHICS STATEMENT

This work collects no new data and recruits no human subjects. The benchmarks are public graphs with anonymized features. The supplementary code reproduces the reported evaluation and gives no capability for identifying individuals in a live system.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Richard Bergna, Sergio Calvo Ordoñez, Felix Opolka, Pietro Liò, and José Miguel Hernández-Lobato. Uncertainty modeling in graph neural networks via stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7 (1):108–116, 1995.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018.

Ben Chamberlain, James Rowbottom, Maria I Gorinova, Michael Bronstein, Stefan Webb, and Emanuele Rossi. Grand: Graph neural diffusion. In *International conference on machine learning*, pp. 1407–1418. PMLR, 2021.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Society, 1997.

Clemens Damke and Eyke Hüllermeier. Linear opinion pooling for uncertainty quantification on graphs. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024. URL https://openreview.net/forum?id=qLGkfpXTSn.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

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

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Combining neural networks with personalized pagerank for classification on graphs. In *International Conference on Learning Representations*, 2019.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

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

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

David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and computational harmonic analysis*, 30(2):129–150, 2011.

Arman Hasanzadeh, Ehsan Hajiramezanali, Krishna Narayanan, Nick Duffield, Mingyuan Zhou, and Xiaoning Qian. Semi-implicit graph variational auto-encoders. *Advances in neural information processing systems*, 32, 2019.

Arman Hasanzadeh, Ehsan Hajiramezanali, Shahin Boluki, Mingyuan Zhou, Nick Duffield, Krishna Narayanan, and Xiaoning Qian. Bayesian graph neural networks with adaptive connection sampling. In *International conference on machine learning*, pp. 4094–4104. PMLR, 2020.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50. IEEE, 2019.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Hans Hao-Hsun Hsu, Yuesong Shen, Christian Tomani, and Daniel Cremers. What makes graph neural networks miscalibrated? *Advances in Neural Information Processing Systems*, 35:13775–13786, 2022.

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.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36:26699–26721, 2023.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362, 2017.

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

Jintang Li, Ruofan Wu, Wangbin Sun, Liang Chen, Sheng Tian, Liang Zhu, Changhua Meng, Zibin Zheng, and Weiqiang Wang. What's behind the mask: Understanding masked graph modeling for graph autoencoders. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1268–1279, 2023.

Ningyi Liao, Zihao Yu, Siqiang Luo, and Gao Cong. Hubgt: Fast graph transformer with decoupled hierarchy labeling. *Advances in Neural Information Processing Systems*, 38:22346–22376, 2026.

Jenny Liu, Aviral Kumar, Jimmy Ba, Jamie Kiros, and Kevin Swersky. Graph normalizing flows. *Advances in Neural Information Processing Systems*, 32, 2019.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Sitao Luan, Chenqing Hua, Minkai Xu, Qincheng Lu, Jiaqi Zhu, Xiao-Wen Chang, Jie Fu, Jure Leskovec, and Doina Precup. When do graph neural networks help with node classification? investigating the homophily principle on node distinguishability. *Advances in Neural Information Processing Systems*, 36:28748–28760, 2023.

Kaushalya Madhawa, Katushiko Ishiguro, Kosuke Nakago, and Motoki Abe. Graphnvp: An invertible flow model for generating molecular graphs. *arXiv preprint arXiv:1905.11600*, 2019.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, volume 8, pp. 1, 2012.

Yin Cheng Ng, Nicolò Colombo, and Ricardo Silva. Bayesian semi-supervised learning with graph gaussian processes. *Advances in Neural Information Processing Systems*, 31, 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Meng Qu, Yoshua Bengio, and Jian Tang. Gmnn: Graph markov neural networks. In *International conference on machine learning*, pp. 5241–5250. PMLR, 2019.

Meng Qu, Huiyu Cai, and Jian Tang. Neural structured prediction for inductive node classification. In *International Conference on Learning Representations*, 2022.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

Donald J Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, 1987.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

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

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *International Conference on Learning Representations*, 2020.

Hamed Shirzad, Ameya Velingker, Balaji Venkatachalam, Danica J Sutherland, and Ali Kemal Sinop. Exphormer: Sparse transformers for graphs. In *International conference on machine learning*, pp. 31613–31632. PMLR, 2023.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pp. 5–42, 2011.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. Gold: Graph out-of-distribution detection via implicit adversarial latent generation. In *International Conference on Learning Representations*, volume 2025, pp. 45297–45326, 2025.

Haibo Wang, Chuan Zhou, Xin Chen, Jia Wu, Shirui Pan, and Jilong Wang. Graph stochastic neural networks for semi-supervised learning. *Advances in Neural Information Processing Systems*, 33: 19839–19848, 2020.

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.

Qitian Wu, Wentao Zhao, Zenan Li, David P Wipf, and Junchi Yan. Nodeformer: A scalable graph structure learning transformer for node classification. *Advances in neural information processing systems*, 35:27387–27401, 2022.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023a.

Qitian Wu, Chenxiao Yang, Wentao Zhao, Yixuan He, David Wipf, and Junchi Yan. DIFFormer: Scalable (graph) transformers induced by energy constrained diffusion. In *The Eleventh International Conference on Learning Representations*, 2023b.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pp. 5453–5462. pmlr, 2018.

Shenzhi Yang, Bin Liang, An Liu, Lin Gui, Xingkai Yao, and Xiaofang Zhang. Bounded and uniform energy-based out-of-distribution detection for graphs. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=mjh7AOWozN.

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

Jaemin Yoo, Hyunsik Jeon, Jinhong Jung, and U Kang. Accurate node feature estimation with structured variational graph autoencoder. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2336–2346, 2022.

Soroush H Zargarbashi, Simone Antonelli, and Aleksandar Bojchevski. Conformal prediction sets for graph neural networks. In *International Conference on Machine Learning*, pp. 12292–12318. PMLR, 2023.

Yingxue Zhang, Soumyasundar Pal, Mark Coates, and Deniz Ustebay. Bayesian graph convolutional neural networks for semi-supervised classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5829–5836, 2019.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. Uncertainty aware semi-supervised learning on graph data. *Advances in neural information processing systems*, 33:12827–12836, 2020.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 DATASETS

Table 3 lists every graph with its undirected edge count. Cora, CiteSeer, and PubMed carry bag-of-words or TF-IDF features (Sen et al., 2008; Yang et al., 2016; Namata et al., 2012), the Amazon and Coauthor graphs carry bag-of-words and keyword features (Shchur et al., 2018), and WikiCS carries GloVe embeddings of the article text (Mernyei & Cangea, 2020). The loader row-normalises the three citation and two co-authorship graphs. The Amazon graphs keep their binary counts, whose row sums vary so widely that normalising them causes the spectral baselines to fail, and WikiCS keeps its embeddings. The scale channel of Section 3.2 arises on the row-normalised graphs. ogbn-arxiv (Hu et al., 2020) serves only the temporal shift of Appendix B.1.

A.2 BASELINES

Baseline Methods. We compare against eleven baselines. GCN (Kipf & Welling, 2017), GraphSAGE with mean aggregation (Hamilton et al., 2017), GAT (Velickovic et al., 2018), and APPNP (Gasteiger et al., 2019) are two-layer networks of width 64 that train with dropout 0.5, Adam at lr 10^{-2} and weight decay 5×10^{-4}, and early-stopping patience of 50 epochs, except for GAT's reference dropout of 0.6 on features and attention. Their inputs are first projected to width 64 when they have more than 256 features. NodeFormer (Wu et al., 2022) uses two layers of width 64 with four heads and temperature $\tau = 0.25$, and DIFFormer-s (Wu et al., 2023b) uses four single-head layers of width 64 with residual mixing weight 0.5. Exphormer (Shirzad et al., 2023) uses four layers with two heads, attention dropout 0.8, and random expanders of degree 6. We take its hidden dimension from the official per-dataset configuration where one exists and select it on validation accuracy otherwise. GRAND (Chamberlain et al., 2021) keeps the official per-dataset hyperparameters, except that the released weight decay causes training on Coauthor CS to fail under our splits, so we lower it there to 10^{-4}. LGNSDE (Bergna et al., 2025) and MaskGAE (Li et al., 2023) use the authors' released model code, and we read out LGNSDE by averaging ten sampled passes and MaskGAE by a linear probe on its frozen embeddings. The Graph Posterior Network (Stadler et al., 2021) trains under the UCE loss with gradient clipping and keeps its reference defaults on every graph, since its published settings do not cover all of our benchmarks.

Detection Controls and DCE. The deterministic GCN and GAT controls of Section 5.3 keep the width, optimiser, weight decay, and early stopping of the two-layer baselines above, and we score them by the softmax entropy of one evaluation-mode pass. GAT spreads its width over eight heads of width eight, so the two backbones match to within one hundred and sixty parameters on every graph. The label-smoothing control retrains that pair with smoothing 0.1, and the input-normalisation control inserts a linear projection to width 64 followed by LayerNorm before the convolutions. DCE is GCNF's context track trained alone (Appendix A.3), and GAT-deep (LN), trained in the same harness, recovers most of its accuracy lead over a two-layer GCN (Figures 2b and 4d). Every control runs for ten trials per graph on the shared per-trial splits.

Comparison Asymmetries. Table 2 is not matched in capacity or recipe. DCE trains 0.17M to 1.20M parameters and GCNF 1.3M to 2.6M (Table 12), both more than the two-layer width-64 baselines. Both also train with Adam at 10^{-3}, cosine warm restarts, label smoothing 0.1, gradient clipping, and a patience counted in evaluations rather than epochs (Appendix A.3), whereas the baselines use Adam at 10^{-2} with none of these. GAT-deep (LN) quantifies how much of the gap this recipe and capacity close by themselves, and Appendix E.3 reports the under-confidence that label smoothing leaves in GCNF.

A.3 DCE AND GCNF CONFIGURATION

GCNF. On the eight benchmarks GCNF uses flow dimension $d = 32$, context width $d_h = 128$, $K = 4$ dual-track layers with $M = 4$ attention heads, coupling-MLP width 256, $k = 64$ retained eigenmodes, and dropout 0.5 on every graph. The input map is a linear projection ($L_{\mathrm{pre}} = 0$), the gate of the positional encoding starts at $\alpha = e^{-2}$, and the objective is Eq. (9) with $\lambda_z = 1$ and

Table 3: Statistics of the nine graphs in our experiments.

Dataset	Nodes	Edges	Feat.	Cls.	Domain
Cora	2,708	5,429	1,433	7	Citation
CiteSeer	3,327	4,732	3,703	6	Citation
PubMed	19,717	44,338	500	3	Citation
Amazon Photo	7,650	119,081	745	8	Co-purchase
Amazon Computers	13,752	245,861	767	10	Co-purchase
Coauthor CS	18,333	81,894	6,805	15	Co-authorship
Coauthor Physics	34,493	247,962	8,415	5	Co-authorship
WikiCS	11,701	216,123	300	10	Web
ogbn-arxiv	169,343	1,157,799	128	40	Citation

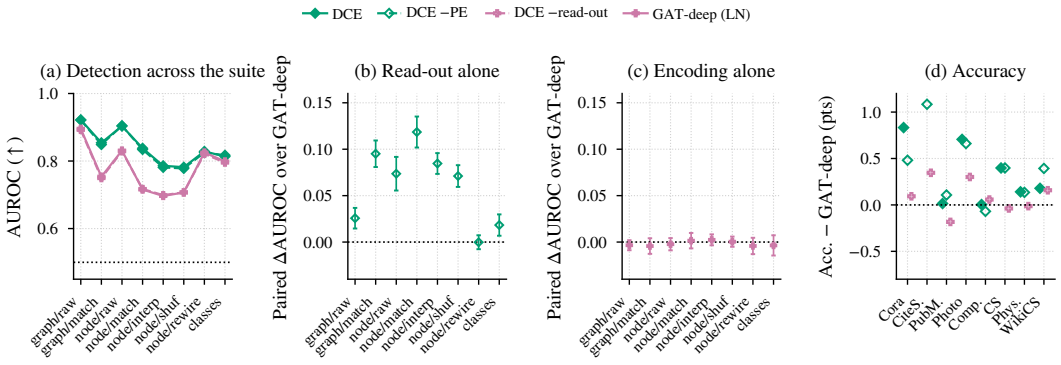

Figure 4: DCE's two distinguishing parts removed one at a time. GAT-deep (LN) lacks both. (a) Mean detection per protocol. (b) Gain from the read-out of $H^{(0)}$ over GAT-deep (LN), without the positional encoding. (c) Gain from the positional encoding over GAT-deep (LN), without the read-out. (d) Accuracy over GAT-deep (LN) per graph. Gaps are paired over 80 trial-graph cells with 95% intervals, and Figure 2d gives both gains inside DCE.

label smoothing $\epsilon = 0.1$ on both cross-entropies. The optimiser is Adam at lr 10^{-3} and weight decay 5×10^{-4} with gradient clipping at 1.0, under cosine annealing with warm restarts ($T_0 = 50$, $T_{\text{mult}} = 2$, $\eta_{\min} = 10^{-5}$), for up to 500 epochs. Early stopping evaluates every 5 epochs with a patience of 50 evaluations, so a run can continue up to 250 epochs past its best checkpoint.

DCE. DCE is GCNF's context track with the latent, the flow, and the auxiliary head removed. It keeps every setting above verbatim and receives no separate tuning.

Prior Smoothness. The smoothness β is trained through its logarithm from $\beta = 1$. Eq. (9) has no likelihood term, so its only gradient reaches β through the sampled noise, scaled by the damping of the retained modes, which is of order $\beta\lambda_i \approx 10^{-2}$ because their eigenvalues lie below 0.1 on most graphs. The smoothness therefore stays within about 2% of its initial value in all eighty runs. This reflects the truncation more than the objective, since at full rank the same objective changes β by 2% to 5% from its initial value of 5 (Appendix F).

Temporal Arm. On ogbn-arxiv the context track of both models replaces graph attention with mean-aggregation SAGE convolutions (Hamilton et al., 2017) to lower memory, and the spectral rank drops to $k = 32$. Training runs for at most 300 epochs with a patience of 40 evaluations, and all other settings are unchanged.

A.4 Evaluation Protocols and Metrics

Shared Protocol. Each trial draws a fresh stratified 60/20/20 split and initialises every model from the trial index, so all methods see the same partition and every per-trial contrast is paired. Checkpoints are selected on the validation split. Dispersions report the population standard deviation over trials (ddof $= 0$) unless a caption says otherwise. Where we test DCE or GCNF against the strongest observed competitor, a post-hoc choice, the paired t-test is descriptive, and we correct each family of eight comparisons by Holm's method (Holm, 1979). Protocol-level contrasts pool the 80 trial-graph cells of a protocol into one paired t-test and correct across the eight protocols of a comparison.

Table 4: Secondary detection metrics under the protocols and checkpoints of Table 1, aggregated as there, with the false-positive rate at 95% TPR above and the risk-coverage area below, both lower-is-better. **Red bold** is best and blue underline second best.

False-positive rate at 95% TPR

Method	graph/raw	graph/matched	node/raw	node/matched	node/interp	node/shuffle	node/rewire	classes
GCN	0.964 ± 0.054	0.849 ± 0.047	0.980 ± 0.026	0.911 ± 0.042	0.915 ± 0.041	0.920 ± 0.030	0.522 ± 0.142	0.573 ± 0.172
GAT	0.912 ± 0.138	0.760 ± 0.065	0.967 ± 0.036	0.871 ± 0.049	0.877 ± 0.044	0.884 ± 0.039	0.456 ± 0.196	0.492 ± 0.168
GCN (LN)	0.819 ± 0.058	0.841 ± 0.046	0.908 ± 0.035	0.906 ± 0.042	0.911 ± 0.035	0.912 ± 0.031	0.476 ± 0.173	0.599 ± 0.152
GAT (LN)	0.727 ± 0.162	0.758 ± 0.074	0.886 ± 0.055	0.870 ± 0.049	0.871 ± 0.043	0.877 ± 0.043	0.418 ± 0.192	0.511 ± 0.163
MC-Dropout (GCN)	0.911 ± 0.108	0.855 ± 0.047	0.960 ± 0.037	0.915 ± 0.039	0.917 ± 0.039	0.920 ± 0.031	0.509 ± 0.138	0.578 ± 0.169
MC-Dropout (GAT)	0.461 ± 0.271	0.847 ± 0.025	0.866 ± 0.087	0.917 ± 0.027	0.921 ± 0.025	0.919 ± 0.025	**0.400 ± 0.184**	0.588 ± 0.157
Deep Ensemble	0.946 ± 0.061	0.843 ± 0.055	0.971 ± 0.028	0.908 ± 0.047	0.911 ± 0.043	0.917 ± 0.034	0.505 ± 0.151	0.560 ± 0.170
GNNSafe	0.945 ± 0.090	0.790 ± 0.060	0.982 ± 0.022	0.924 ± 0.025	0.927 ± 0.020	0.932 ± 0.018	0.471 ± 0.144	**0.467 ± 0.190**
GNNSafe (LN)	0.741 ± 0.130	0.789 ± 0.060	0.923 ± 0.024	0.920 ± 0.022	0.920 ± 0.022	0.924 ± 0.020	0.441 ± 0.156	0.510 ± 0.196
GPN	0.909 ± 0.028	0.831 ± 0.053	0.916 ± 0.021	0.870 ± 0.031	0.876 ± 0.032	0.887 ± 0.027	0.703 ± 0.128	0.589 ± 0.144
GAT-deep (LN)	0.355 ± 0.151	0.694 ± 0.184	0.499 ± 0.203	0.732 ± 0.173	0.829 ± 0.093	0.819 ± 0.078	0.565 ± 0.234	0.669 ± 0.154
DCE (Ours)	0.307 ± 0.189	**0.458 ± 0.210**	0.357 ± 0.182	0.480 ± 0.210	**0.732 ± 0.122**	**0.764 ± 0.075**	0.530 ± 0.196	0.633 ± 0.176
GCNF (DCE + flow)	**0.259 ± 0.149**	0.462 ± 0.211	**0.289 ± 0.132**	**0.477 ± 0.207**	0.749 ± 0.121	0.771 ± 0.079	0.562 ± 0.190	0.642 ± 0.163

Risk-coverage area

Method	graph/raw	graph/matched	node/raw	node/matched	node/interp	node/shuffle	node/rewire	classes
GCN	0.640 ± 0.171	0.369 ± 0.034	0.327 ± 0.112	0.159 ± 0.025	0.162 ± 0.024	0.165 ± 0.022	0.068 ± 0.021	0.470 ± 0.076
GAT	0.640 ± 0.229	0.335 ± 0.028	0.309 ± 0.113	0.150 ± 0.023	0.153 ± 0.021	0.155 ± 0.020	0.059 ± 0.023	0.450 ± 0.072
GCN (LN)	0.355 ± 0.040	0.368 ± 0.032	0.157 ± 0.023	0.158 ± 0.023	0.159 ± 0.022	0.161 ± 0.021	0.062 ± 0.023	0.479 ± 0.069
GAT (LN)	0.323 ± 0.060	0.335 ± 0.032	0.150 ± 0.026	0.149 ± 0.022	0.149 ± 0.021	0.151 ± 0.022	0.056 ± 0.022	0.455 ± 0.075
MC-Dropout (GCN)	0.483 ± 0.126	0.375 ± 0.035	0.231 ± 0.070	0.163 ± 0.025	0.164 ± 0.023	0.166 ± 0.021	0.066 ± 0.019	0.471 ± 0.076
MC-Dropout (GAT)	0.252 ± 0.086	0.383 ± 0.026	0.168 ± 0.052	0.171 ± 0.020	0.171 ± 0.018	0.170 ± 0.018	**0.056 ± 0.020**	0.481 ± 0.080
Deep Ensemble	0.511 ± 0.138	0.367 ± 0.035	0.251 ± 0.085	0.159 ± 0.025	0.161 ± 0.024	0.164 ± 0.022	0.066 ± 0.021	0.467 ± 0.074
GNNSafe	0.664 ± 0.221	0.339 ± 0.030	0.339 ± 0.108	0.173 ± 0.018	0.176 ± 0.018	0.178 ± 0.017	0.071 ± 0.015	**0.444 ± 0.079**
GNNSafe (LN)	0.314 ± 0.046	0.333 ± 0.023	0.172 ± 0.020	0.171 ± 0.015	0.172 ± 0.016	0.174 ± 0.016	0.067 ± 0.016	0.455 ± 0.073
GPN	0.455 ± 0.024	0.396 ± 0.041	0.177 ± 0.010	0.155 ± 0.016	0.157 ± 0.017	0.160 ± 0.019	0.099 ± 0.019	0.477 ± 0.071
GAT-deep (LN)	0.210 ± 0.026	0.307 ± 0.061	0.065 ± 0.022	0.107 ± 0.036	0.120 ± 0.032	0.115 ± 0.029	0.073 ± 0.034	0.504 ± 0.096
DCE (Ours)	0.200 ± 0.035	**0.236 ± 0.048**	0.047 ± 0.016	0.063 ± 0.023	**0.087 ± 0.024**	**0.090 ± 0.023**	0.070 ± 0.032	0.487 ± 0.081
GCNF (DCE + flow)	**0.191 ± 0.027**	0.238 ± 0.053	**0.041 ± 0.011**	**0.063 ± 0.025**	0.089 ± 0.026	0.091 ± 0.024	0.075 ± 0.035	0.489 ± 0.080

Interpretation of Null Results. The rank-64 prior swap has no resolving power on Cora, Coauthor CS, and Coauthor Physics, where its two arms coincide numerically (Appendix F). On the other graphs the minimum detectable paired effect over ten trials is 0.002 to 0.030 AUROC-H at $\alpha = 0.05$ and power 0.8, dominated by PubMed. The full-rank family resolves 0.004 to 0.006 AUROC-H against the variance-matched control and up to 0.019 for the neighbour-conditioned arms. The cells of Appendix G.1 are bounded by their prespecified equivalence regions of ± 0.005 accuracy, ± 0.01 AUROC-H and AURC-H, and ± 0.01 nats of pairwise likelihood. Every null claim should be read as bounded by these values.

Sampled Scores. MC-Dropout keeps dropout active over $S = 30$ passes, and over GAT it also keeps the attention-coefficient dropout of every head active. The deep ensemble averages the softmax outputs of five independently trained GCNs. Every sampled method, GCNF included, estimates mutual information as the entropy of the mean prediction minus the mean entropy of the individual predictions.

Detection Metrics. We write AUROC-H and AURC-H wherever the ranking score is the predictive entropy, which every GCNF-against-GCNF contrast uses. AURC retains nodes in order of increasing score and reads the risk at each coverage as the fraction of anomalous nodes retained, so its floor is not zero. When the two blocks are balanced, as at graph scope, a perfect detector scores $(1 - \log 2)/2 \approx 0.153$ in the large-sample limit against 0.5 in expectation for a random order, and at node scope it scores $0.2 + 0.8 \log 0.8 \approx 0.021$ against 0.2. Tied scores retain the order returned by the sort, which can favour either block. GNNSafe (Wu et al., 2023a) propagates the logit energy E_v of Section 3.1 over two steps of

$$E \leftarrow \alpha_{\text{prop}} E + (1 - \alpha_{\text{prop}})\, D^{-1} A E \tag{10}$$

with $\alpha_{\text{prop}} = 0.5$ and no auxiliary OOD exposure.

B The Node-Level Protocol Suite

Relation to Published Protocols. In GPN's reference implementation the clean bag-of-words features are row-normalised while the Gaussian draws are not (Stadler et al., 2021), which opens the scale channel of Section 3.2. GNNSafe scores its generated graph in a separate forward pass against the original (Wu et al., 2023a), which is the graph scope. Our raw graph-scope setting combines both designs.

Construction Details. The shuffle leaves fewer than one anomalous node per trial-graph cell in place. Rewiring removes every edge incident to $\mathcal{O}$ before it draws the new ones. Edges drawn by

other nodes of $\mathcal{O}$ can attach to a node, so degrees are kept exactly for most of $\mathcal{O}$ on the citation and co-authorship graphs but for only about a quarter on the denser Amazon graphs and WikiCS. The held-out third of the classes is rounded and redrawn on every trial, its nodes continue to participate in message passing, and each model is retrained on the reduced label space. Raw noise multiplies the mean row norm by only 1.8 on the Amazon graphs and 6.2 on WikiCS, and matched noise keeps it within 8% of its clean value on every graph.

Detection Scores. Every node-scope protocol and the semantic one score anomalous and clean nodes in one forward pass, so a detector never sees a clean counterpart of an anomalous node. We train each backbone once per graph and trial and read it by entropy and by every further score it supports, namely propagated energy on the GCN, DCE, GAT-deep (LN), and GCNF backbones; mutual information for the sampled methods; total evidence for GPN; and the exact latent likelihood for GCNF. Tables 1 and 7 hold the resulting twenty-three scores, in which MC-Dropout and GNNSafe are read-outs of the same trained GCN or GAT that supplies an entropy row. LGNSDE (Bergna et al., 2025) needs thirty passes of an adaptive SDE solver per protocol, so we score it only on Cora, CiteSeer, and Amazon Photo under the seven perturbation protocols. The entropy of its path mean tracks plain GCN throughout and inverts with it under raw noise, since its input GCN also reads raw features. Its mutual information trails that entropy except under raw noise, and DCE outperforms it on every feature protocol and matches it under rewiring.

Paired Margins. Figure 2a gives the three paired margins of Section 5.2. DCE's margin over GAT-deep (LN) survives Holm correction on every feature protocol. It is smallest under raw graph-level noise, where the entire corrupted pass is already distinct for both encoders, and largest under matched noise, whose rows match real features in scale and moments but differ in content, which the read-out of $H^{(0)}$ exposes before any aggregation. It vanishes under rewiring and is small but significant on held-out classes. The margin over the best dedicated score, chosen post hoc among all scores of MC-Dropout, the ensemble, GPN, and GNNSafe, is widest at node scope, where a two-layer backbone averages each anomalous node with its clean neighbours.

Component Analysis of DCE. Two further arms remove the positional encoding or the read-out of $H^{(0)}$ from DCE under the same recipe, splits, and draws, so that with DCE and GAT-deep (LN) they complete a 2×2 design (Figure 4). Figure 2d gives what each part adds inside DCE, and panels (b) and (c) give what it adds to GAT-deep (LN) when the other part is absent. Removing the read-out costs detection on every feature protocol and on held-out classes after Holm correction, but not under rewiring, and it leaves an arm indistinguishable from GAT-deep (LN) under every protocol. The read-out adds the same margin with or without the positional encoding, most under matched noise at either scope and substantially under shuffling. The encoding changes no protocol in either setting after correction and leaves accuracy essentially unchanged (panel d).

Reported Tables. Table 4 gives the false-positive rate at 95% true-positive rate and the risk-coverage area behind Table 1. Tables 5 and 6 break the four protocols that separate the methods down by benchmark, and Table 7 gives the remaining scores.

B.1 Real Distribution Shift on ogbn-arxiv

Training sees only the papers published up to 2017, split $70/10/20$ into training, validation, and the in-distribution pool, and every later paper is scored in the same pass (Hu et al., 2020; Wu et al., 2023a). GCNF and DCE share the SAGE context track of Appendix A.3. No score reliably identifies the newer papers (Table 8). The default scores stay close to chance in every year, and both GNNSafe variants invert. GCNF's latent likelihood, the one score that averages higher at 0.575 ± 0.076, varies too much across trials to count as detection. GCNF classifies the papers of 2018 slightly better than its in-distribution pool and loses little accuracy by 2020, and DCE is at least as accurate, so the shift has little effect on the quantities these scores measure.

B.2 The Deterministic Control Ladder

We train the ladder of Section 5.3 in the mechanism-study harness and score it under raw graph-level noise, so its means differ from the rows of Table 1 by at most 0.004 (Figure 5). Plain GCN and GAT fall below chance on the five row-normalised graphs, and on exactly those graphs plain GCN's entropy on corrupted nodes falls below its entropy on clean ones, whereas GCNF's rises above it

Table 5: Per-benchmark detection AUROC (↑) under graph-level raw corruption and leave-out classes, reported as mean±population s.d. Values near 0 are inverted-sign scores, read $1 - \text{AUROC}$. **Red bold** is best and blue underline second best.

Method	Cora	CiteSeer	PubMed	Amazon Photo	Amazon Computers	CS	Physics	WikiCS
Graph-scope raw corruption								
GCN	0.109 ± 0.009	0.044 ± 0.004	0.093 ± 0.009	0.620 ± 0.004	0.619 ± 0.012	0.249 ± 0.015	0.349 ± 0.034	0.549 ± 0.009
GAT	0.085 ± 0.013	0.024 ± 0.004	0.059 ± 0.008	0.769 ± 0.044	0.763 ± 0.064	0.105 ± 0.021	0.115 ± 0.024	0.537 ± 0.067
GCN (LN)	0.760 ± 0.018	0.705 ± 0.045	0.652 ± 0.017	0.648 ± 0.010	0.620 ± 0.010	0.753 ± 0.006	0.742 ± 0.007	0.633 ± 0.003
GAT (LN)	0.771 ± 0.018	0.688 ± 0.059	0.680 ± 0.035	0.705 ± 0.149	0.675 ± 0.090	0.811 ± 0.017	0.897 ± 0.019	0.713 ± 0.013
MC-Dropout (GCN)	0.496 ± 0.039	0.311 ± 0.086	0.310 ± 0.027	0.614 ± 0.004	0.628 ± 0.010	0.742 ± 0.020	0.652 ± 0.048	0.625 ± 0.011
MC-Dropout (GAT)	0.911 ± 0.041	0.814 ± 0.030	0.614 ± 0.023	0.825 ± 0.061	0.795 ± 0.072	0.984 ± 0.005	**0.972 ± 0.024**	0.790 ± 0.038
Deep Ensemble	0.532 ± 0.028	0.330 ± 0.064	0.183 ± 0.019	0.632 ± 0.005	0.634 ± 0.007	0.756 ± 0.019	0.586 ± 0.041	0.586 ± 0.012
GNNSafe	0.004 ± 0.003	0.000 ± 0.000	0.020 ± 0.011	0.687 ± 0.014	0.703 ± 0.020	0.002 ± 0.001	0.126 ± 0.053	0.612 ± 0.016
GNNSafe (LN)	0.757 ± 0.029	0.717 ± 0.019	0.665 ± 0.011	0.754 ± 0.016	0.728 ± 0.023	0.845 ± 0.018	0.829 ± 0.013	0.690 ± 0.006
GPN	0.529 ± 0.006	0.525 ± 0.009	0.544 ± 0.005	0.542 ± 0.012	0.590 ± 0.077	0.534 ± 0.005	0.547 ± 0.006	0.579 ± 0.005
GAT-deep (LN)	0.847 ± 0.040	0.827 ± 0.020	**0.882 ± 0.015**	0.886 ± 0.020	0.899 ± 0.009	0.987 ± 0.004	0.950 ± 0.007	0.888 ± 0.014
DCE (ours)	**0.946 ± 0.009**	**0.877 ± 0.026**	0.822 ± 0.021	0.943 ± 0.006	0.947 ± 0.009	**0.994 ± 0.001**	0.942 ± 0.006	0.905 ± 0.012
GCNF (DCE + flow)	0.934 ± 0.016	0.868 ± 0.024	0.867 ± 0.031	**0.951 ± 0.007**	**0.965 ± 0.008**	0.993 ± 0.000	0.945 ± 0.009	**0.959 ± 0.009**
Leave-out classes								
GCN	0.823 ± 0.049	0.794 ± 0.020	0.684 ± 0.082	0.865 ± 0.044	0.805 ± 0.069	0.904 ± 0.033	0.935 ± 0.010	0.790 ± 0.025
GAT	0.852 ± 0.041	0.811 ± 0.029	**0.719 ± 0.059**	0.868 ± 0.102	0.797 ± 0.076	0.922 ± 0.029	0.951 ± 0.008	0.844 ± 0.026
GCN (LN)	0.810 ± 0.040	0.779 ± 0.029	0.698 ± 0.086	0.850 ± 0.048	0.790 ± 0.058	0.886 ± 0.035	0.926 ± 0.016	0.770 ± 0.024
GAT (LN)	0.837 ± 0.047	0.808 ± 0.025	0.681 ± 0.080	0.895 ± 0.035	0.821 ± 0.053	0.913 ± 0.028	0.937 ± 0.015	0.839 ± 0.028
MC-Dropout (GCN)	0.820 ± 0.052	0.792 ± 0.022	0.681 ± 0.082	0.858 ± 0.041	0.793 ± 0.072	0.898 ± 0.034	0.935 ± 0.010	0.795 ± 0.026
MC-Dropout (GAT)	0.793 ± 0.055	0.763 ± 0.052	0.638 ± 0.071	0.803 ± 0.097	0.773 ± 0.054	0.880 ± 0.036	0.923 ± 0.017	0.809 ± 0.047
Deep Ensemble	0.812 ± 0.057	0.807 ± 0.023	0.689 ± 0.086	0.876 ± 0.028	0.813 ± 0.067	0.908 ± 0.032	0.935 ± 0.011	0.795 ± 0.026
GNNSafe	**0.869 ± 0.046**	**0.815 ± 0.032**	0.674 ± 0.106	0.920 ± 0.043	**0.841 ± 0.063**	**0.936 ± 0.036**	**0.969 ± 0.003**	**0.858 ± 0.028**
GNNSafe (LN)	0.832 ± 0.069	0.794 ± 0.065	0.696 ± 0.098	**0.922 ± 0.044**	0.820 ± 0.057	0.929 ± 0.043	0.963 ± 0.007	0.826 ± 0.039
GPN	0.777 ± 0.064	0.795 ± 0.025	0.681 ± 0.083	0.827 ± 0.049	0.784 ± 0.063	0.882 ± 0.040	0.923 ± 0.034	0.819 ± 0.060
GAT-deep (LN)	0.806 ± 0.054	0.747 ± 0.049	0.559 ± 0.068	0.907 ± 0.020	0.781 ± 0.104	0.899 ± 0.031	0.881 ± 0.035	0.811 ± 0.020
DCE (ours)	0.797 ± 0.044	0.757 ± 0.025	0.629 ± 0.080	0.913 ± 0.025	0.798 ± 0.082	0.912 ± 0.019	0.884 ± 0.017	0.816 ± 0.015
GCNF (DCE + flow)	0.802 ± 0.036	0.760 ± 0.022	0.631 ± 0.079	0.900 ± 0.019	0.803 ± 0.081	0.914 ± 0.022	0.874 ± 0.012	0.803 ± 0.034

Table 6: Per-benchmark node-level detection AUROC (↑) with a fifth of the test split made anomalous and the graph scored in one pass. **Red bold** is best and blue underline second best.

Method	Cora	CiteSeer	PubMed	Amazon Photo	Amazon Computers	CS	Physics	WikiCS
Node-level matched corruption								
GCN	0.632 ± 0.028	0.704 ± 0.021	0.636 ± 0.012	0.543 ± 0.024	0.543 ± 0.013	0.605 ± 0.009	0.567 ± 0.009	0.550 ± 0.013
GAT	0.639 ± 0.020	0.693 ± 0.017	0.620 ± 0.008	0.545 ± 0.022	0.538 ± 0.015	0.596 ± 0.006	0.573 ± 0.011	0.539 ± 0.011
GCN (LN)	0.640 ± 0.026	0.693 ± 0.024	0.628 ± 0.011	0.544 ± 0.018	0.554 ± 0.014	0.598 ± 0.012	0.562 ± 0.010	0.552 ± 0.013
GAT (LN)	0.637 ± 0.021	0.689 ± 0.024	0.620 ± 0.013	0.552 ± 0.019	0.543 ± 0.016	0.596 ± 0.010	0.582 ± 0.012	0.536 ± 0.012
MC-Dropout (GCN)	0.632 ± 0.031	0.700 ± 0.019	0.635 ± 0.012	0.532 ± 0.024	0.535 ± 0.015	0.601 ± 0.009	0.561 ± 0.009	0.546 ± 0.014
MC-Dropout (GAT)	0.607 ± 0.028	0.653 ± 0.016	0.585 ± 0.009	0.520 ± 0.025	0.521 ± 0.018	0.564 ± 0.008	0.535 ± 0.009	0.520 ± 0.013
Deep Ensemble	0.646 ± 0.021	0.708 ± 0.026	0.637 ± 0.012	0.544 ± 0.023	0.545 ± 0.015	0.605 ± 0.007	0.567 ± 0.009	0.552 ± 0.013
GNNSafe	0.583 ± 0.021	0.651 ± 0.019	0.555 ± 0.007	0.519 ± 0.022	0.520 ± 0.014	0.587 ± 0.009	0.550 ± 0.009	0.520 ± 0.019
GNNSafe (LN)	0.585 ± 0.024	0.647 ± 0.018	0.561 ± 0.013	0.532 ± 0.021	0.536 ± 0.013	0.586 ± 0.014	0.550 ± 0.010	0.519 ± 0.017
GPN	0.571 ± 0.014	0.638 ± 0.023	0.612 ± 0.011	0.546 ± 0.033	0.529 ± 0.021	0.555 ± 0.011	0.570 ± 0.013	0.570 ± 0.014
GAT-deep (LN)	0.599 ± 0.063	0.607 ± 0.021	0.705 ± 0.014	0.691 ± 0.039	0.646 ± 0.022	0.899 ± 0.025	0.809 ± 0.039	0.769 ± 0.015
DCE (ours)	0.763 ± 0.032	0.804 ± 0.023	**0.729 ± 0.011**	0.887 ± 0.014	0.812 ± 0.025	0.969 ± 0.005	**0.931 ± 0.007**	**0.809 ± 0.012**
GCNF (DCE + flow)	**0.784 ± 0.025**	**0.831 ± 0.024**	0.703 ± 0.025	**0.889 ± 0.020**	**0.827 ± 0.017**	**0.973 ± 0.005**	0.927 ± 0.006	0.795 ± 0.015
Structural rewiring								
GCN	0.797 ± 0.029	0.821 ± 0.021	0.687 ± 0.014	0.904 ± 0.023	0.810 ± 0.015	0.906 ± 0.008	0.871 ± 0.017	0.758 ± 0.018
GAT	0.829 ± 0.024	0.796 ± 0.025	0.719 ± 0.013	**0.960 ± 0.009**	0.892 ± 0.024	0.922 ± 0.005	0.892 ± 0.015	0.868 ± 0.018
GCN (LN)	**0.850 ± 0.028**	**0.833 ± 0.018**	0.701 ± 0.011	0.899 ± 0.018	0.785 ± 0.017	**0.955 ± 0.007**	0.917 ± 0.009	0.763 ± 0.033
GAT (LN)	0.844 ± 0.015	0.805 ± 0.029	0.727 ± 0.012	0.960 ± 0.005	0.884 ± 0.025	0.944 ± 0.007	**0.922 ± 0.009**	**0.879 ± 0.015**
MC-Dropout (GCN)	0.804 ± 0.028	0.821 ± 0.023	0.700 ± 0.013	0.902 ± 0.023	0.795 ± 0.015	0.908 ± 0.007	0.865 ± 0.017	0.764 ± 0.019
MC-Dropout (GAT)	0.827 ± 0.023	0.789 ± 0.028	0.717 ± 0.011	0.952 ± 0.007	0.869 ± 0.030	0.928 ± 0.004	0.884 ± 0.018	0.870 ± 0.018
Deep Ensemble	0.818 ± 0.023	0.826 ± 0.019	0.689 ± 0.011	0.917 ± 0.012	0.806 ± 0.019	0.913 ± 0.006	0.869 ± 0.007	0.762 ± 0.014
GNNSafe	0.731 ± 0.025	0.725 ± 0.029	**0.745 ± 0.008**	0.825 ± 0.019	0.837 ± 0.017	0.833 ± 0.011	0.786 ± 0.014	0.848 ± 0.010
GNNSafe (LN)	0.767 ± 0.034	0.729 ± 0.023	0.745 ± 0.018	0.838 ± 0.024	0.817 ± 0.020	0.885 ± 0.015	0.847 ± 0.015	0.856 ± 0.015
GPN	0.701 ± 0.029	0.707 ± 0.038	0.709 ± 0.009	0.764 ± 0.071	0.755 ± 0.041	0.679 ± 0.029	0.847 ± 0.035	0.665 ± 0.019
GAT-deep (LN)	0.805 ± 0.042	0.764 ± 0.031	0.636 ± 0.017	0.957 ± 0.007	**0.915 ± 0.021**	0.857 ± 0.010	0.832 ± 0.036	0.836 ± 0.020
DCE (ours)	0.809 ± 0.035	0.766 ± 0.021	0.637 ± 0.016	0.943 ± 0.009	0.911 ± 0.006	0.867 ± 0.015	0.849 ± 0.042	0.846 ± 0.014
GCNF (DCE + flow)	0.805 ± 0.022	0.753 ± 0.020	0.617 ± 0.033	0.931 ± 0.013	0.898 ± 0.018	0.861 ± 0.014	0.840 ± 0.024	0.820 ± 0.014

everywhere (panel b). Label smoothing, which caps the confidence of the training targets, deepens the inversion on these five graphs (Section 5.3), so overconfident targets are not its cause. The projection with LayerNorm raises both backbones well above chance, and DCE closes most of the remaining gap to GCNF without any latent. GCNF's base-mode and sampled scores agree closely on every graph (panel c), so its remaining improvement over DCE is almost entirely independent of sampling.

C Theoretical Analysis

This appendix proves Proposition 6.1 and Theorem 6.2 and states and proves the budget decomposition and the local-decoder certificate. Notation follows Sections 3.1 and 6, and every proof starts from the covariance of the spectral noise.

Lemma C.1 (Covariance of the spectral noise)**.** *The d channels of $\boldsymbol{z}_{\mathrm{sp}}$ in Eq.* (5) *are independent, and each has covariance*

$$\Sigma = I - U_k \mathrm{diag}(\boldsymbol{w})\, U_k^\top, \qquad w_i = \frac{\beta\lambda_i}{1+\beta\lambda_i} \in (0,1). \tag{11}$$

Table 7: Detection AUROC (↑) of the ten scores absent from Table 1, whose thirteen rows complete the twenty-three, as eight-benchmark means of ten-trial means. Spread and density scores lose their raw-noise lead under matched noise, while propagated energy and evidence gain on held-out classes.

Score	graph/raw	graph/matched	node/raw	node/matched	node/interp	node/shuffle	node/rewire	classes
DCE, propagated energy	0.893	0.809	0.775	0.714	0.682	0.683	0.755	0.848
GAT-deep (LN), propagated energy	0.890	0.722	0.728	0.631	0.624	0.632	0.763	0.819
GCNF, propagated energy	0.906	0.808	0.795	0.720	0.682	0.683	0.750	0.851
GCNF, base-mode entropy	0.935	0.852	0.924	0.841	0.782	0.780	0.816	0.811
GCNF, mutual information	0.865	0.710	0.857	0.708	0.667	0.688	0.702	0.750
GCNF, latent likelihood	0.739	0.482	0.731	0.492	0.448	0.521	0.539	0.576
MC-Dropout (GCN), mutual information	0.753	0.624	0.660	0.570	0.575	0.578	0.840	0.760
MC-Dropout (GAT), mutual information	0.946	0.491	0.684	0.495	0.508	0.521	0.728	0.482
Deep Ensemble, mutual information	0.706	0.654	0.615	0.581	0.573	0.576	0.817	0.799
GPN, total evidence	0.667	0.639	0.584	0.565	0.525	0.493	0.430	0.855

Row v therefore has law $\mathcal{N}(\mathbf{0}, \sigma_v^2 I_d)$ with $\sigma_v^2 = 1 - \sum_{i \le k} w_i u_{vi}^2$, and for any two nodes $u \neq v$

$$\mathbb{E}\big[\|\boldsymbol{n}_u - \boldsymbol{n}_v\|^2\big] = d\Big(2 - \sum_{i=1}^{k} w_i \,(u_{ui} - u_{vi})^2\Big) > 0, \tag{12}$$

which is non-increasing in β for every pair and strictly decreasing for every pair whose retained spectral coordinates differ.

Proof. Each channel applies the symmetric map $T = I + U_k \mathrm{diag}(\boldsymbol{\delta}) U_k^\top$ to standard Gaussian noise, and the orthonormality of U_k gives

$$TT^\top = I + U_k \mathrm{diag}(2\boldsymbol{\delta} + \boldsymbol{\delta}^2)\, U_k^\top. \tag{13}$$

Since $(1+\delta_i)^2 = (1+\beta\lambda_i)^{-1} = 1 - w_i$, this is Σ. The row law and σ_v^2 follow from its diagonal, and

$$\mathbb{E}\,\|\boldsymbol{n}_u - \boldsymbol{n}_v\|^2 = d\,(\Sigma_{uu} + \Sigma_{vv} - 2\Sigma_{uv}) \tag{14}$$

expands to the stated form. Every retained eigenvalue is strictly positive, so $\partial w_i / \partial \beta = \lambda_i / (1 + \beta\lambda_i)^2 > 0$ gives the monotonicity. Positivity holds because every $w_i < 1$ and

$$\sum_{i=1}^{k} (u_{ui} - u_{vi})^2 = (e_u - e_v)^\top U_k U_k^\top (e_u - e_v) \le \|e_u - e_v\|^2 = 2. \tag{15}$$

□

Proof of Proposition 6.1. We argue conditionally on the auxiliary randomness ξ, which is independent of N. Each conclusion below is an identity that holds for ξ-almost every realisation, so taking $\mathbb{E}_\xi$ of both sides returns the unconditional statement, and this step covers the training-time dropout masks.

GCNF meets Eq. (3) at every ξ. Its context track involves no Z, since $H^{(0)}$ depends only on X and the cached eigenvectors and each layer of Eq. (2) reads only $H^{(\ell-1)}$. At depth zero $\boldsymbol{z}_v^{(0)} = \phi_z(X)_v + \boldsymbol{n}_v$, and the coupling of Eq. (37) updates $\boldsymbol{z}_{b,v}$ through $q_v^{(\ell)} = [\boldsymbol{z}_{a,v}^{(\ell-1)} \| H_v^{(\ell)} \| H_v^{(0)}]$ alone, so by induction $\boldsymbol{z}_v^{(\ell)}$ is a measurable function of $\boldsymbol{n}_v$ at every depth. The head reads $[\boldsymbol{z}_v^{(K)} \| H_v^{(K)} \| H_v^{(0)}]$, so the logits at v equal $G_v(\boldsymbol{n}_v)$ for a measurable G_v fixed by X, $\mathcal{G}$, and the weights.

The predictive law at v is then the pushforward of the row marginal P_v under G_v, so equal row marginals give equal per-node predictive laws. Consider a separable risk with integrable terms,

$$R_{\boldsymbol{\theta}}(N) = \frac{1}{|\mathcal{V}_{\mathrm{tr}}|} \sum_{v \in \mathcal{V}_{\mathrm{tr}}} \ell_{v,\boldsymbol{\theta}}(\boldsymbol{n}_v). \tag{16}$$

Linearity of expectation gives

$$\mathbb{E}_P R_{\boldsymbol{\theta}} = \frac{1}{|\mathcal{V}_{\mathrm{tr}}|} \sum_{v \in \mathcal{V}_{\mathrm{tr}}} \mathbb{E}_{P_v} \ell_{v,\boldsymbol{\theta}} = \mathbb{E}_{P'} R_{\boldsymbol{\theta}}. \tag{17}$$

Equality as functions of $\boldsymbol{\theta}$ on an open domain gives equal expected gradients wherever differentiation and expectation commute. Nothing fixes the cross-node terms $\mathbb{E}[\nabla \ell_u \nabla \ell_v^\top]$, so the gradient covariances and the joint output laws may differ. For the spectral prior, Lemma C.1 gives the row law $\mathcal{N}(\mathbf{0}, \sigma_v^2 I_d)$, and $w_i < 1$ gives $|\sigma_v^2 - 1| \le m_v$. Orthonormality gives $\sum_v m_v = \|U_k\|_F^2 = k$. Since m_v is a diagonal entry of the projection $U_k U_k^\top$, it lies in $[0,1]$, whence $\sigma_v^2 \ge 1 - \max_i w_i > 0$. □

Table 8: Temporal shift on ogbn-arxiv, where models train on papers up to 2017 and score 2018 onwards in the same pass. The last three columns give AUROC against each publication year.

Method	AUROC (↑)	FPR@95 (↓)	AUPR (↑)	2018	2019	2020
GCN	0.541 ± 0.002	0.904 ± 0.002	0.828 ± 0.002	0.532	0.543	0.564
GCN (LN)	0.540 ± 0.003	0.910 ± 0.005	0.829 ± 0.001	0.532	0.541	0.563
MC-Dropout (GCN)	0.517 ± 0.002	0.917 ± 0.002	0.817 ± 0.002	0.512	0.516	0.536
Deep Ensemble	0.540 ± 0.001	0.905 ± 0.002	0.828 ± 0.001	0.531	0.542	0.562
GNNSafe	0.385 ± 0.001	0.981 ± 0.001	0.763 ± 0.001	0.395	0.376	0.387
GNNSafe (LN)	0.369 ± 0.002	0.985 ± 0.001	0.759 ± 0.001	0.382	0.358	0.370
DCE (ours)	0.505 ± 0.003	0.938 ± 0.001	0.811 ± 0.002	0.498	0.507	0.519
GCNF (DCE + flow)	0.505 ± 0.001	0.941 ± 0.003	0.811 ± 0.001	0.499	0.507	0.520

Remark C.2 (Scope of the hypothesis)**.** *Proposition 6.1 fails without the node-wise condition. Take two adjacent nodes with $d = 2$ and compare independent standard-normal rows with the degenerate prior $\boldsymbol{n}_u = \boldsymbol{n}_v$, which has the same marginal $\mathcal{N}(\mathbf{0}, I_2)$ at every node. A coupling that adds the neighbour's untouched channel to the transformed channel at v yields $(n_{v,1}, n_{v,2} + n_{u,1})$, whose channels have covariance zero under the first prior and one under the second, so the read-out $z_1 + z_2$ has variance 3 against 5. Within the node-wise class, the law of every per-node quantity (such as the predictive entropy, the latent mutual information, or a fixed nonconformity score), and hence expected accuracy, is invariant to the correlation shape. Quantities that rank nodes against one another, such as detection AUROC and selective risk, depend on the joint law and lie outside the proposition, as do a data-dependent calibration quantile and joint coverage.*

Proposition C.3 (Budget, profile, and shape decomposition)**.** *Let w_i and σ_v^2 be as in Lemma C.1. Under node-wise decoding the predictive law at every node depends on the prior only through*

$$\sigma_v^2 = B + \pi_v, \qquad B := 1 - \frac{1}{n}\sum_{i=1}^{k} w_i, \qquad |\pi_v| \;\le\; w_{\max}\Big(m_v + \frac{k}{n}\Big), \tag{18}$$

where the budget B *is the average marginal variance, the* profile π_v *is its node-wise deviation, and the cross-node correlation* shape *contributes nothing. The log-determinant of Eq.* (40) *satisfies*

$$\frac{\partial}{\partial \log \beta}\Big(-\tfrac{1}{n}\log|\det J_{\mathrm{sp}}|\Big) \;=\; \frac{d}{2}\big(1 - B\big). \tag{19}$$

The decomposition and $1 - B \le k/n$ hold for any damping spectral filter with per-mode variances in $(0, 1]$ and any choice of k retained modes.

Proof. By Eq. (3) the predictive law at v is the pushforward of the row marginal $\mathcal{N}(\mathbf{0}, \sigma_v^2 I_d)$ of Lemma C.1, so it depends on the prior only through σ_v^2. Averaging with $\sum_v u_{vi}^2 = 1$ gives $n^{-1}\sum_v \sigma_v^2 = B$, and $\pi_v = -\sum_i w_i(u_{vi}^2 - 1/n)$ obeys the stated bound by the triangle inequality. For the GMRF, Eq. (40) gives

$$-\frac{1}{n}\log|\det J_{\mathrm{sp}}| = \frac{d}{2n}\sum_{i=1}^{k}\log(1 + \beta\lambda_i), \tag{20}$$

whose derivative in $\log\beta$ is

$$\frac{d}{2n}\sum_{i=1}^{k}\frac{\beta\lambda_i}{1+\beta\lambda_i} = \frac{d}{2n}\sum_{i=1}^{k} w_i = \frac{d}{2}(1 - B). \tag{21}$$

A filter with per-mode variances $g(\lambda_i) \in (0, 1]$ keeps the decomposition, the profile bound, and $1 - B \le k/n$ with $w_i = 1 - g(\lambda_i)$. Its log-determinant derivative instead depends on how g varies with β and need not equal $\frac{d}{2}(1 - B)$. The heat filter $g(\lambda) = e^{-\beta\lambda}$, for instance, gives

$$\frac{d}{2n}\sum_{i=1}^{k}\beta\lambda_i \;>\; \frac{d}{2n}\sum_{i=1}^{k}\big(1 - e^{-\beta\lambda_i}\big) \tag{22}$$

whenever some $\beta\lambda_i > 0$. □

A variance bound does not bound the output laws, since a decoder may be arbitrarily sensitive to the scale of what it reads.

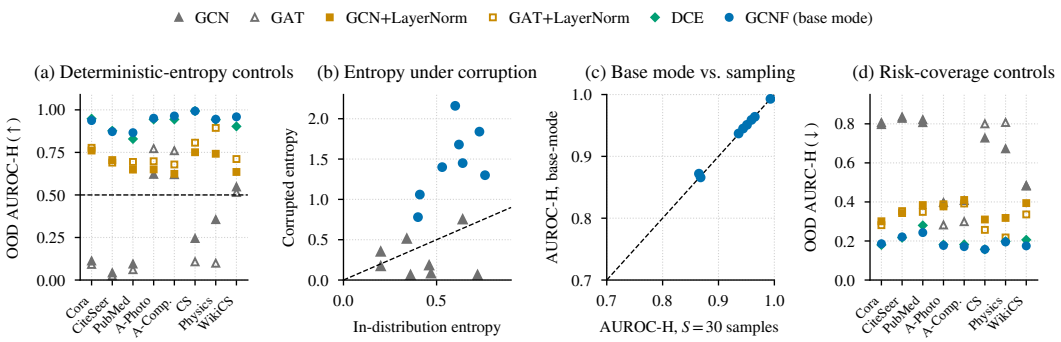

Figure 5: The control ladder under raw graph-level noise. (a, d) Detection and risk-coverage from plain backbones to DCE and GCNF. (b) Predictive entropy, clean against corrupted. (c) Base-mode against S=30 sampled scores.

Proof of Theorem 6.2. The argument also yields the data-dependent certificate

$$\frac{1}{n}\sum_v \mathrm{TV}(\mu_v^{(k)}, \mu_v^{(0)}) \le \frac{1}{2n}\sum_v \sqrt{d[-a_v - \log(1-a_v)]} \tag{23}$$

and shows that the k/n rate is attained, since for every fixed $\rho \in (0,1)$ some orthonormal rank-k basis and node-wise decoder have an average total-variation gap of a positive constant times k/n.

Row marginals. By Lemma C.1, row v of the structured noise has law $P_v = \mathcal{N}(\mathbf{0}, (1-a_v)I_d)$ with $a_v = \sum_i w_i u_{vi}^2$, and the isotropic control has $Q_v = \mathcal{N}(\mathbf{0}, I_d)$. The leverage score m_v lies in $[0,1]$, so $0 \le a_v \le \rho\, m_v \le \rho < 1$ and both laws are nondegenerate.

Contraction through the decoder. The log density ratio

$$\log \frac{dP_v}{dQ_v}(x) = -\frac{d}{2}\log(1-a_v) - \frac{a_v}{2(1-a_v)}\,\|x\|^2 \tag{24}$$

and $\mathbb{E}_{P_v}\|x\|^2 = d(1-a_v)$ give

$$D_{\mathrm{KL}}(P_v \| Q_v) = \frac{d}{2}[-a_v - \log(1-a_v)]. \tag{25}$$

By Eq. (3) the output laws are the pushforwards $\mu_v^{(k)} = G_{v\#}P_v$ and $\mu_v^{(0)} = G_{v\#}Q_v$. With $r = dP_v/dQ_v$ and $Y = G_v(X)$ for $X \sim Q_v$, the likelihood ratio of the output laws at Y is $\mathbb{E}_{Q_v}[r(X) \mid Y]$, and conditional Jensen for $t \mapsto t\log t$ gives

$$D_{\mathrm{KL}}(\mu_v^{(k)} \| \mu_v^{(0)}) = \mathbb{E}_{Q_v}[\mathbb{E}[r \mid Y]\log \mathbb{E}[r \mid Y]] \le \mathbb{E}_{Q_v}[r \log r] = D_{\mathrm{KL}}(P_v \| Q_v). \tag{26}$$

This data-processing step uses only the measurability of G_v, so no Lipschitz or Jacobian condition enters, and with Eq. (25) it proves Eq. (6). Pinsker's inequality gives

$$\mathrm{TV}\big(\mu_v^{(k)}, \mu_v^{(0)}\big) \le \frac{1}{2}\sqrt{d\,[-a_v - \log(1-a_v)]}, \tag{27}$$

and averaging proves Eq. (23).

Reduction to rank. For $0 \le a < 1$ the logarithmic series gives

$$-a - \log(1-a) = \sum_{r\ge 2}\frac{a^r}{r} \le \frac{a^2}{2(1-a)}, \tag{28}$$

so $\mathrm{TV}(\mu_v^{(k)}, \mu_v^{(0)}) \le a_v\sqrt{d/(8(1-\rho))}$. Orthonormality removes every graph-specific detail from the average, since

$$\frac{1}{n}\sum_v a_v = \frac{1}{n}\sum_{i=1}^k w_i \sum_v u_{vi}^2 = \frac{1}{n}\sum_{i=1}^k w_i = 1 - B \le \rho\,\frac{k}{n}, \tag{29}$$

Table 9: Twenty-two graph models with a stochastic latent, sorted by what the prediction at v may read of it and by whether the objective is separable, where r is coupling depth or solver steps.

Method	Sampled object	Structure enters via	Read set S_v	Objective	Thm. 6.2
Structured prior on a per-node latent					
GCNF (ours)	latent row $\boldsymbol{n}_v$	spectral sampler	$\{v\}$	separable	yes
SVGA (Yoo et al., 2022)	latent row	objective (GMRF KL)	$\{v\}$	joint	dec. only
Graph GP (Ng et al., 2018)	node function f_v	posterior inference	$\{v\}$	joint	dec. only
Latent draws mixed across nodes					
GNF (Liu et al., 2019)	latent row	message-passing couplings	r-hop ball	separable	no
GraphNVP (Madhawa et al., 2019)	whole latent	relational couplings	all nodes	joint	no
GraphAF (Shi et al., 2020)	per-atom, sequential	autoregressive couplings	preceding nodes	joint	no
LGNSDE (Bergna et al., 2025)	Brownian path	graph-convolution drift	r-hop ball	ELBO	no
SIG-VAE (Hasanzadeh et al., 2019)	row plus shared noise	hierarchical sharing	node pairs	joint	no
The sampled object is shared across nodes					
BGCN (Zhang et al., 2019)	the graph	posterior over graphs	all nodes	joint	no
BBGDC (Hasanzadeh et al., 2020)	edge and weight masks	Bayesian masks	all nodes	ELBO	no
S-BGCN-T-K (Zhao et al., 2020)	dropout weight masks	objective (kernel Dirichlet prior)	all nodes	joint	no
GSNN (Wang et al., 2020)	global function latent	amortised joint posterior	all nodes	joint	no
NodeFormer (Wu et al., 2022)	propagation edges	Gumbel topology	all nodes	separable	no
MC-Dropout (Gal & Ghahramani, 2016)	weight masks	backbone only	all nodes	separable	no
Deep Ensembles (Lakshminarayanan et al., 2017)	parameter draws	backbone only	all nodes	separable	no
Joint label field					
GMNN (Qu et al., 2019)	label variables	pairwise label potentials	label neighbourhood	joint	no
SPN (Qu et al., 2022)	label variables	structured potentials	label neighbourhood	joint	no
Per-node latent, no structured prior					
VGAE (Kipf & Welling, 2016)	latent row	none (isotropic)	node pairs	joint	no
Graph2Gauss (Bojchevski & Günnemann, 2018)	per-node Gaussian	none	node pairs	joint	no
No stochastic latent					
GPN (Stadler et al., 2021)	none	evidence propagation	n/a	separable	n/a
GNNSafe (Wu et al., 2023a)	none	energy propagation	n/a	n/a	n/a
CF-GNN (Huang et al., 2023)	none	conformal correction	n/a	n/a	n/a

which proves Eq. (7). For any measurable h_v with values in $[0,1]$ the variational characterisation of total variation gives

$$\left|\mathbb{E}_{\mu_v^{(k)}} h_v - \mathbb{E}_{\mu_v^{(0)}} h_v\right| \le \mathrm{TV}\big(\mu_v^{(k)}, \mu_v^{(0)}\big), \tag{30}$$

and summing proves the observable claim.

Tightness. Fix $\rho \in (0,1)$ and take $U_k = [e_1, \dots, e_k]$ with every $w_i = \rho$, so $a_v = \rho$ for $v \le k$ and $a_v = 0$ otherwise. A normalised Laplacian realises this basis whenever $k \le n-2$. Take k isolated vertices and a complete graph on the remaining $m = n - k \ge 2$ vertices. Under our isolated-node convention, the former have eigenvalue 1, while the complete component contributes one zero mode and the eigenvalue $m/(m-1) > 1$ with multiplicity $m-1$. The k smallest positive eigenvalues therefore all equal 1, and their eigenvectors span the same subspace as $e_1, \dots, e_k$, so every orthonormal choice of them gives the same a_v, and $\beta = \rho/(1-\rho)$ sets every $w_i = \rho$. Let $c_{\rho,d} := \mathrm{TV}(\mathcal{N}(\mathbf{0}, (1-\rho)I_d), \mathcal{N}(\mathbf{0}, I_d)) > 0$. The binary node-wise decoder that indicates where the likelihood ratio of the two row laws exceeds one keeps this gap on the first k nodes and has gap zero elsewhere, so its average gap is exactly

$$\frac{1}{n}\sum_{v=1}^{n} \mathrm{TV}\big(\mu_v^{(k)}, \mu_v^{(0)}\big) = c_{\rho,d}\,\frac{k}{n}. \tag{31}$$

□

A decoder that reads several latent rows leaves the node-wise class, and Proposition C.4 covers read sets of any shape, such as r-hop couplings and solver trajectories.

Proposition C.4 (Local-decoder certificate)**.** *Fix the graph, features, and decoder weights, and suppose the logits at v are $G_v(N_{S_v})$ for a measurable G_v and a read set $S_v \subseteq \mathcal{V}$, where N_{S_v} collects the latent rows indexed by S_v. Let $\kappa := \max_{u \in \mathcal{V}} |\{v : u \in S_v\}|$ be the largest number of read sets that any one row enters. With Σ the covariance of Lemma C.1, $A_v := \Sigma_{S_v S_v} - I$, and a_v, B, ρ as in*

Theorem 6.2,

$$D_{\mathrm{KL}}\Big(\mu_v^{(k)}\Big\|\mu_v^{(0)}\Big) \le \frac{d}{2}\left\{\operatorname{tr} A_v - \log\det(I+A_v)\right\}, \tag{32}$$

$$\frac{1}{n}\sum_{v=1}^{n}\mathrm{TV}\Big(\mu_v^{(k)},\mu_v^{(0)}\Big) \le \sqrt{\frac{d}{8(1-\rho)}}\,\frac{1}{n}\sum_{v=1}^{n}\|A_v\|_F$$

$$\le \kappa\sqrt{\frac{d}{8(1-\rho)}}\,(1-B) \;\le\; \rho\,\kappa\sqrt{\frac{d}{8(1-\rho)}}\,\frac{k}{n}. \tag{33}$$

At $S_v=\{v\}$, where $\kappa=1$, every line reduces to Theorem 6.2.

Proof. By Lemma C.1 the read-set law is $\mathcal{N}(\mathbf{0},\Sigma_{S_vS_v}\otimes I_d)$ under the rank-k prior and $\mathcal{N}(\mathbf{0},I\otimes I_d)$ under isotropic noise. Write $A_v=-B_v$ with $B_v:=U_{S_v}\mathrm{diag}(\boldsymbol{w})U_{S_v}^\top$, where U_{S_v} is the row submatrix of U_k. Then $B_v\preceq\rho\,U_{S_v}U_{S_v}^\top$, and $U_{S_v}U_{S_v}^\top$ is a principal submatrix of the projection $U_kU_k^\top$, so Cauchy interlacing puts the spectrum of B_v in $[0,\rho]$ and that of $\Sigma_{S_vS_v}$ in $[1-\rho,1]$. Data processing through the fixed measurable G_v bounds the output divergence by the divergence of the two read-set laws, which the Gaussian formula gives as

$$\frac{d}{2}\big\{\operatorname{tr}\Sigma_{S_vS_v}-|S_v|-\log\det\Sigma_{S_vS_v}\big\}, \tag{34}$$

and this is Eq. (32). For $\alpha\in[-\rho,0]$ the logarithmic series gives $\alpha-\log(1+\alpha)\le\alpha^2/(2(1-\rho))$, so summing over the eigenvalues of A_v yields

$$D_{\mathrm{KL}}\big(\mu_v^{(k)}\,\|\,\mu_v^{(0)}\big)\le\frac{d\,\|A_v\|_F^2}{4(1-\rho)}, \tag{35}$$

and Pinsker's inequality gives the first bound of Eq. (33). Since B_v is positive semidefinite, $\|A_v\|_F\le\operatorname{tr}B_v=\sum_{u\in S_v}a_u$. A sum over v counts each row once for every read set it enters, so

$$\sum_{v=1}^{n}\sum_{u\in S_v}a_u\le\kappa\sum_{u=1}^{n}a_u=\kappa\,n(1-B), \tag{36}$$

which is the second bound, and the third follows from $\sum_u a_u=\sum_i w_i\le\rho k$. At $S_v=\{v\}$ the matrix A_v is the scalar $-a_v$, so the chain reduces to that of the theorem. □

The factor κ cannot be dropped. On the rank-one basis $U_k=e_1$ of the tightness construction, read sets that all contain row 1 give $\kappa=n$, and a decoder that reads row 1 keeps an average gap of $c_{\rho,d}$ at every n while $1-B=\rho/n$. Beyond singleton read sets the row marginals no longer determine the output laws, since the degenerate prior of Remark C.2 has the row marginals of white noise but a decoder with $\kappa=2$ separates it. A bound that covers every Gaussian prior must therefore read more of Σ than its diagonal, as Eq. (32) does. The last two lines of Eq. (33) read the diagonal alone only because the degenerate prior lies outside the damping class, inside which every entry of B_v obeys $(B_v)_{uu'}^2\le a_ua_{u'}$. Appendix D evaluates Eq. (32) on the benchmarks.

D Which Graph Models the Hypothesis Covers

Table 9 sorts twenty-two graph models with a stochastic or distributional latent by what they sample, how graph structure enters its law, which latent rows the prediction at v may read, and whether the objective scores nodes one at a time. The last two are the hypotheses of Proposition 6.1. Theorem 6.2 needs only the read set, whereas the risk and gradient clauses of the proposition also need a separable objective, so the final column records whether the theorem applies to a model as stated, to its decoder alone ("dec. only"), or not at all.

Three of the twenty-two decode node-wise in the sense of Eq. (3), and only GCNF also trains separably, so the strict class, in which the results constrain training as well as prediction, contains no published model we could find. The other two keep a node-wise decoder and let the prior act elsewhere. SVGA (Yoo et al., 2022) reads its GMRF latent through the per-node decoder $g(\boldsymbol{z}_v)=W\boldsymbol{z}_v+\boldsymbol{b}$ but carries the Laplacian quadratic form $\mathrm{tr}(U^\top KU)$ of its structured KL term in the objective, so the predictive

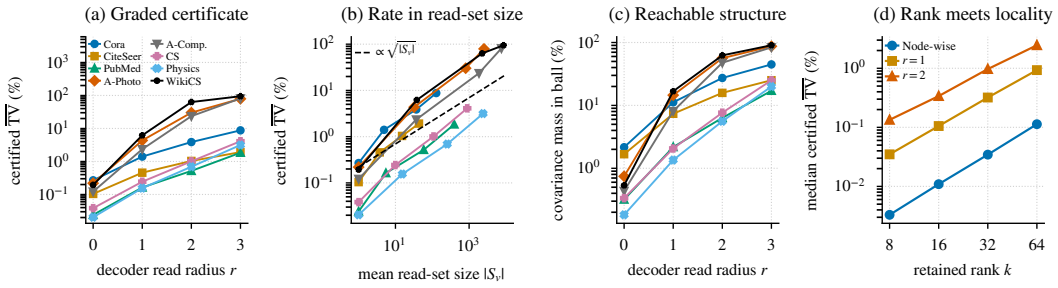

Figure 6: Certificates for decoders that read more than one latent row, at $\beta = 1$, $k = 64$, and $d = 32$. (a) Certificate for r-hop read sets. (b) Certificate against read-set size. (c) Covariance mass in the ball. (d) Rank sweep.

claim of Proposition 6.1 binds it while the risk and gradient claims do not. The graph Gaussian process of Ng et al. (2018) decodes each node through its own likelihood but conditions a joint posterior, so the prior's covariance arrives through inference. Zhao et al. (2020) likewise place their graph-kernel Dirichlet prior in the training term. Models whose structured prior influences what they learn therefore route it through a non-separable objective or an inference step, which Proposition 6.1 leaves open, whereas models that keep a separable objective let draws move between nodes, which Proposition C.4 quantifies.

Graded Certificate for Local Decoders. Figure 6 evaluates Eq. (32) from the cached spectra for balls of radius zero to three, the receptive fields of a coupling flow of that depth or of a latent SDE with that many solver steps, and radius zero reproduces Table 10. The certificate tracks how many rows a decoder reads rather than how far it reaches. One hop keeps it at a few percent on every graph, whereas two hops raise it above one half on WikiCS, since a two-hop ball there and on the co-purchase graphs holds hundreds to thousands of rows. Beyond one hop it grows roughly as the square root of the mean read-set size (panel b) and stays far below the linear envelope of Proposition C.4, which bounds every read set by the most frequently read row. On Cora a single node of degree 168 sets $\kappa = 169$ for one-hop read sets and so inflates the envelope by two orders of magnitude, whereas Eq. (32) on the same read sets keeps the certificate near 1.4% (panel a). Most of a row's prior covariance also lies outside its one-hop ball (panel c), so a one-hop decoder can reach only a minor part of the prior's cross-node structure. The certificate holds at fixed weights, so a model trained from scratch under a joint objective lies outside it. The rank sweep of panel (d) shows that the premises interact, since cutting the rank eightfold shrinks the median certificate about twice as much under node-wise decoding as under a two-hop read. A full-rank prior read by a local decoder is therefore the setting this analysis constrains least, and Appendix F tests it by retraining.

E GCNF: Architecture, Objective, and Likelihood

GCNF is the instrument of Section 6, and Algorithm 1 gives one training step and the inference rule.

Coupling Layers. Each coupling splits $\boldsymbol{z}^{(\ell-1)}$ into an identity half $\boldsymbol{z}_a$ and a transformed half $\boldsymbol{z}_b$, whose roles alternate between layers, and conditions on the DCE states of Eqs. (1) and (2),

$$q_v^{(\ell)} = \big[\boldsymbol{z}_{a,v}^{(\ell-1)} \| H_v^{(\ell)} \| H_v^{(0)}\big], \qquad \boldsymbol{z}_{b,v}^{(\ell)} = \boldsymbol{z}_{b,v}^{(\ell-1)} \cdot e^{2\tanh(\mathrm{MLP}_s(q_v^{(\ell)}))} + \mathrm{MLP}_t(q_v^{(\ell)}), \tag{37}$$

so the update at v reads row v of the latent and no other.

Auxiliary Probe. The auxiliary term of Eq. (9) applies the main cross-entropy to a linear probe of $\boldsymbol{z}_v^{(K)}$, discarded at inference, which pushes the latent to carry class information. The feature-derived latent location alone satisfies it, and nothing in the objective asks predictions to vary across draws or rewards correlation between nodes.

Identifiability Under the Objective. GCNF meets both hypotheses of Proposition 6.1, since its decoder is node-wise and Eq. (9) collects the main and auxiliary losses at node v into one per-node term $\ell_{v,\boldsymbol{\theta}}$. For any joint noise law P the expected objective therefore factorises over the rows $\boldsymbol{n}_v$ of

N,

$$\mathbb{E}_{N\sim P}[\mathcal{L}(\boldsymbol{\theta}; N)] = |\mathcal{V}_{\mathrm{tr}}|^{-1} \sum_{v\in\mathcal{V}_{\mathrm{tr}}} \mathbb{E}_{\boldsymbol{n}_v\sim P_v}[\ell_{v,\boldsymbol{\theta}}(\boldsymbol{n}_v)], \tag{38}$$

and equal-marginal priors share this objective and its expected gradient whenever the per-node gradients are integrable. This factorisation leaves no incentive for correlation shape but does not rule out every training-time effect, since a joint draw can still change the stochastic-gradient covariance and β acts on the marginal-variance profile.

Incentive Against Using the Draw. Each training step evaluates Eq. (9) on a single draw. Fix the network, take a finite set of draws with positive weights ω_s summing to one, and write $p_s(c)$ for the probability that the head assigns to class c under draw s. For label-smoothed targets $q_c > 0$ the strict concavity of the logarithm gives

$$\sum_s \omega_s \sum_c q_c\big(-\log p_s(c)\big) \;\geq\; \sum_c q_c\Big(-\log \sum_s \omega_s\, p_s(c)\Big), \tag{39}$$

with equality exactly when no $p_s(c)$ varies with s. Among networks with the same draw-averaged prediction, the objective is therefore lowest for one whose prediction does not vary between draws, whatever the joint law of the noise, and the auxiliary probe meets the same inequality. The argument needs only that the draw is independent of the labels, so it applies equally to MC-Dropout, whose network also trains on one mask per step and whose one-hot target penalises the variation of the true-class probability alone. This is the log-loss counterpart of the regularising effect of training with noise (Bishop, 1995). Separability removes any reward for correlation shape at given marginals, and Eq. (39) penalises the draw's influence on each node, so the pairwise score of Appendix G.1 is the one term in this paper that rewards that influence, and only when an edge shares it.

Latent Likelihood Monitoring. We train the flow by Eq. (9) alone. Adding $-\log p(Z^{(K)} \mid X, \mathcal{G})$ would fit no observed density, since the network that a gradient step updates generates $Z^{(K)}$ itself, so the term would only reward the model for placing its own samples where its base measure is dense. In a five-trial sweep on a wider variant of the flow ($d = 64$ behind a two-layer graph-convolutional input map), the term costs accuracy on every benchmark at weight 0.1, and smaller weights change accuracy in both directions without raising the eight-benchmark mean.

Base-Mode Evaluation and Uncertainty. At evaluation we zero the noise term of Eq. (8), which leaves the mode $\boldsymbol{z}_\star^{(0)} = \phi_z(X)$ of the conditional base, so a point prediction is one deterministic pass. We do not call this a MAP rule, because an invertible nonlinear $f_{\boldsymbol{\theta}}$ need not carry modes to modes. Uncertainty scores draw $S = 30$ samples from the surrogate $\tilde{p}_0^{(k)}$ and read the predictive entropy of the mean softmax and the mutual information between prediction and latent draw (Houlsby et al., 2011; Gal et al., 2017), which carries no weight-space disagreement, unlike its ensemble counterpart. Appendix E.3 compares the two read-outs.

Component Ablation. Retraining GCNF without the coupling's condition on the DCE states, without the copy of $H^{(0)}$ inside that condition, without the auxiliary head, or without the positional encoding changes accuracy by at most 0.29 points on any benchmark, within the trial spread of the full model. Every removal but the positional encoding leaves the path from the DCE states to the head intact, which is consistent with the small effect of each removal.

E.1 The Latent Log-Likelihood

Both sources of volume change in GCNF are closed-form. The spectral sampler of Eq. (5) contributes

$$\log|\det J_{\mathrm{sp}}| = d\sum_{i=1}^{k}\log(1+\delta_i) = -\tfrac{d}{2}\sum_{i=1}^{k}\log(1+\beta\lambda_i), \tag{40}$$

and the triangular Jacobian of coupling layer ℓ contributes the sum of its scales (Kingma & Dhariwal, 2018),

$$\log|\det J^{(\ell)}| = \sum_{v\in\mathcal{V}}\sum_{j=1}^{d/2} s_{v,j}^{(\ell)}. \tag{41}$$

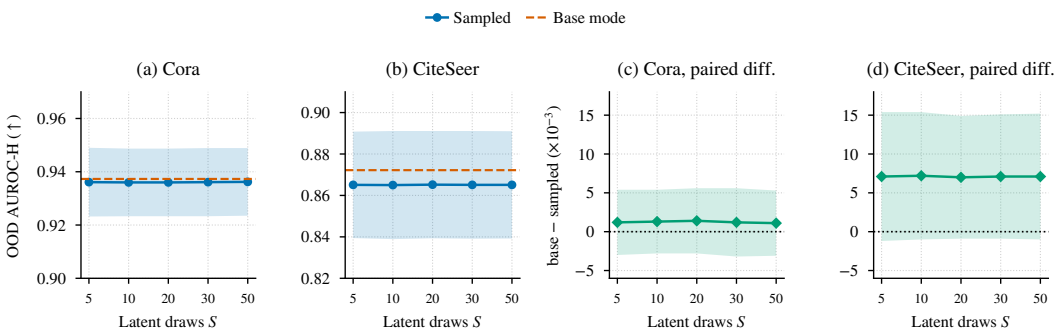

Figure 7: Sampling sensitivity on Cora and CiteSeer. (a, b) Mean OOD AUROC-H against the number of draws S, base-mode dashed. (c, d) Paired base-mode minus S-sample difference, ± 1 s.d. bands.

Algorithm 1 GCNF: One Training Step and Inference

Require: features X, graph $\mathcal{G}$, cached eigenpairs $(U_k, \{\lambda_i\})$, labelled nodes $\mathcal{V}_{\text{tr}}$

$H^{(0)} \leftarrow \mathrm{LN}(XW_h) + e^{\gamma}\,\rho(U_k)$ {DCE input, Eq. (1)}
draw $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I)$; $\boldsymbol{z}_{\text{sp}} \leftarrow \boldsymbol{\varepsilon} + U_k \mathrm{diag}(\boldsymbol{\delta})\, U_k^{\top} \boldsymbol{\varepsilon}$ {surrogate $\tilde{p}_0^{(k)}$, Eq. (5)}
$\boldsymbol{z}^{(0)} \leftarrow \mathrm{LN}(XW_z) + \boldsymbol{z}_{\text{sp}}$ {Eq. (8)}
$\log p \leftarrow \log \mathcal{N}(\boldsymbol{\varepsilon}; \mathbf{0}, I) + \frac{d}{2} \sum_i \log(1 + \beta\lambda_i)$ {Eq. (40)}
for $\ell = 1, \ldots, K$ **do**
  $H^{(\ell)} \leftarrow \mathrm{LN}\big(H^{(\ell-1)} + \mathrm{GATConv}_M(H^{(\ell-1)}, \mathcal{G})\big)$
  split $\boldsymbol{z}^{(\ell-1)} \to (\boldsymbol{z}_a, \boldsymbol{z}_b)$, with halves swapped each layer; $q \leftarrow [\boldsymbol{z}_a \| H^{(\ell)} \| H^{(0)}]$
  $\boldsymbol{z}_b \leftarrow \boldsymbol{z}_b \odot e^{s(q)} + t(q)$; $\boldsymbol{z}^{(\ell)} \leftarrow [\boldsymbol{z}_a \| \boldsymbol{z}_b]$
  $\log p \leftarrow \log p - \sum_j s_j(q)$ {coupling log-det, Eq. (41)}
end for
$\hat{y} \leftarrow \psi\big([\boldsymbol{z}^{(K)} \| H^{(K)} \| H^{(0)}]\big)$
minimise $\mathcal{L}_{\text{CE}} + \lambda_z \mathcal{L}_{\text{aux}}$ {Eq. (9), with $\log p$ monitored rather than optimised}
Inference. Point prediction: one pass with $\boldsymbol{z}_{\text{sp}} = \mathbf{0}$ (base mode). Uncertainty: repeat lines 2 to 11 for S noise draws, average the softmax outputs, and evaluate the predictive entropy and latent mutual information of Section 6.3.

Each coupling is a smooth bijection with the explicit inverse

$$\boldsymbol{z}_{b,v}^{(\ell-1)} = \big(\boldsymbol{z}_{b,v}^{(\ell)} - \mathrm{MLP}_t(q_v^{(\ell)})\big)\, e^{-s_v^{(\ell)}}, \tag{42}$$

and its Jacobian determinant is a product of factors e^s with $s = 2\tanh(\cdot) \in (-2, 2)$. The shift by $\phi_z(X)$ has unit Jacobian and the base is a nondegenerate Gaussian, so the change-of-variables formula (Papamakarios et al., 2021) gives $Z^{(K)}$ a density that is positive everywhere, integrates to one, and has the exact log-likelihood

$$\log p(Z^{(K)} \mid X, \mathcal{G}) = \sum_v \log \mathcal{N}(\boldsymbol{\varepsilon}_v; \mathbf{0}, I) - \log|\det J_{\text{sp}}| - \sum_{\ell=1}^{K} \log|\det J^{(\ell)}|, \tag{43}$$

where $\boldsymbol{\varepsilon}_v$ is the pre-transform Gaussian draw. This likelihood concerns a variable the model generates itself, so it is neither a density fitted to observations nor a posterior, and it guarantees no calibrated label probabilities (Appendix E.3). It makes the latent law and the effect of each prior swap auditable without variational inference.

E.2 Ingredient Ablation

For the ingredient ablation of Section 6.4 we retrain GCNF under the configuration of Appendix A.3 with one change per variant, on the same ten per-trial splits. The three variants replace the spectral base by isotropic noise of the same dimension, set the draw to zero so that the model trains and predicts at $\boldsymbol{z}^{(0)} = \phi_z(X)$, or replace every coupling by the identity so that the draw reaches the head unshaped. Paired t-tests against the full model are Holm-corrected across graphs under raw graph-level noise (Figure 3a and c) and across protocols in the suite (panel d).

Under raw graph-level noise, GCNF's lead over DCE survives correction only on Amazon Computers and WikiCS. The isotropic base changes no graph (Appendix F), and identity couplings reduce detection most on WikiCS and PubMed but barely change it on the co-authorship graphs and Amazon Photo. Removing the draw raises detection on every graph, significantly on Cora and the two co-authorship graphs, and leaves accuracy unchanged. For the suite we rescore the same checkpoints under the seven perturbation protocols with the suite's own node selection and draws, so the full model reproduces the suite's GCNF rows. Identity couplings then hurt only under the two raw protocols and help under every other protocol, whereas removing the draw helps only under those two, leaves matched noise and interpolation unchanged, and hurts shuffled and rewired nodes. Both patterns place the flow's effect in its deterministic path, which improves detection when raw noise places a node's features far outside the data and slightly weakens it for anomalies that remain in range. The isotropic base changes no protocol, and base-mode entropy gives the same verdicts.

E.3 Sampling and Calibration Diagnostics

The sampled read-out adds little to the base-mode pass. Across the eight benchmarks the mean mutual information between prediction and latent draw is at least two orders of magnitude below the mean predictive entropy, so most in-distribution uncertainty is shared across draws, which is the absorption that Eq. (39) rewards. Replacing the thirty-sample predictive mean by the base-mode pass changes the test NLL by at most 0.003 nats and the Brier score by at most 0.0007 on any benchmark. Retraining without the spectral surrogate changes both by even less, which is consistent with Theorem 6.2, although the theorem covers neither retrained weights nor the unbounded log loss. Figure 7 sweeps the number of draws from five to fifty in a post-hoc study that selects no setting. Detection is flat in S, and the base-mode pass lies slightly above the S-sample mean by a gap that does not shrink with more draws, so the gap reflects decoding at the mode rather than Monte Carlo error. We keep $S = 30$ because the mutual information needs a variance estimate.

Label smoothing (Szegedy et al., 2016) leaves GCNF under-confident. Smoothing at ϵ over C classes trains the head towards a target of $1 - \epsilon(C-1)/C$ on the true class, just above 0.9 at $\epsilon = 0.1$ on every benchmark, and GCNF's mean confidence stays below that target on every graph. Its 15-bin expected calibration error (Guo et al., 2017) lies between 0.03 and 0.08. Mean confidence falls below accuracy by up to eight points on the co-authorship and co-purchase graphs, whose accuracy exceeds the target, and differs from it negligibly on CiteSeer, whose accuracy lies below the target. An exactly tractable latent density therefore delivers no binned calibration, and the retrained prior-free variant is equally calibrated, so the prior has no role in it.

F Dissecting the Latent Noise: Shape versus Scale

Realised Rank-64 Intervention. Table 10 gives what the deployed surrogate changes, computed from the cached spectra. The bound $1 - B \le k/n$ of Proposition C.3 is loose here by a factor of six to thirty-seven, because truncation keeps the modes the prior damps least. The retained eigenvalues lie below 0.3 on every graph and below 0.1 on the citation and co-authorship graphs, so at $\beta = 1$ every weight w_i stays below a quarter. Eq. (23) then certifies that no fixed measurable node-wise decoder separates the structured prior from isotropic noise by more than a fraction of a percent in average total variation. On Cora, Coauthor CS, and Coauthor Physics the full and prior-free arms of the ingredient ablation produce identical test accuracies on all ten trials and AUROC-H values that agree far below the trial-to-trial spread. The rank-64 swap is therefore a check on the implementation, and the full-rank experiment carries the evidence.

Full-Rank Experiment. On Cora, WikiCS, and Amazon Computers we sample the GMRF of Eq. (4) over all $n - c$ positive modes of a cached dense spectrum, under the configuration of Appendix A.3 and the same ten per-trial splits. The smoothness starts at $\beta = 5$, far from isotropy, and trains end to end. The positional encoding keeps the first 64 modes in every arm, so only the stochastic base changes rank. The bulk of the spectrum lies near $\lambda \approx 1$, so on Cora at the learned β the marginal variance averages $B \approx 0.23$ and varies across nodes about twenty-eight times as much as at rank 64. By Proposition 6.1 the shape cannot change a per-node law at fixed weights, whereas the budget and the profile can, and three controls vary the profile, the budget, and the decoder's access to the shape. A variance-matched isotropic base keeps the budget B and sets the profile to zero. Rescaling the full-spectrum prior to unit budget keeps its shape and its relative profile σ_v^2/B. Conditioning

Table 10: Realised rank-64 intervention at $\beta = 1$, computed from the cached spectra, where $\overline{\mathrm{TV}}_{\mathrm{cert}}$ is the decoder-independent bound of Eq. (23) at $d = 32$.

Dataset	n	$1 - B$ **(%)**	k/n **(%)**	$\overline{\mathrm{TV}}_{\mathrm{cert}}$ **(%)**	$\mathrm{sd}(\sigma_v^2)$	σ_v^2 **range**
Cora	2,708	0.134	2.36	0.269	0.0028	$[0.964, 1.000]$
CiteSeer	3,327	0.052	1.92	0.104	0.0013	$[0.982, 1.000]$
Amazon Photo	7,650	0.109	0.84	0.220	0.0045	$[0.892, 1.000]$
WikiCS	11,701	0.096	0.55	0.195	0.0055	$[0.881, 1.000]$
Amazon Computers	13,752	0.061	0.47	0.122	0.0030	$[0.910, 1.000]$
Coauthor CS	18,333	0.019	0.35	0.038	0.0013	$[0.968, 1.000]$
PubMed	19,717	0.012	0.32	0.024	0.0008	$[0.976, 1.000]$
Coauthor Physics	34,493	0.010	0.19	0.020	0.0009	$[0.967, 1.000]$

each coupling on its neighbours' identity half exposes the cross-node correlations to the transport at unchanged log-determinant, which deliberately breaks the node-wise hypothesis. None of the three shows a detected difference on any of the three graphs. Read alone, these nulls show only that GCNF does not use the structured channel under a separable objective, not that the channel cannot help when an objective rewards it, and Appendix G.1 supplies that objective.

F.1 Equivalence of the Full-Rank Contrasts

We read the three full-rank contrasts as two one-sided tests (Schuirmann, 1987; Lakens, 2017) against regions of practical equivalence (ROPE) fixed before the analysis, namely ± 0.005 for accuracy and ± 0.01 for AUROC-H and AURC-H, and a cell passes when its paired 90% interval lies inside its region (Figure 8a). Base-mode and sampled accuracy coincide exactly and count once. The four contrast-graph-metric cells that fail after Holm correction all lie on Cora, whose accuracy varies most across trials, and each of their intervals contains zero. The interventions therefore leave the endpoints unchanged, rather than changing them by an amount that ten trials cannot resolve.

F.2 Exact Marginal Matching

The variance-matched control keeps the budget but not the profile, so the last control matches every row marginal exactly and leaves the two arms differing only in the cross-node correlation that Proposition 6.1 rules out at fixed weights. On each of the three graphs we form the full-spectrum factor $T = I + U\mathrm{diag}(\boldsymbol{\delta})U^\top$ at a fixed $\beta = 5$. The structured arm draws $N = T\boldsymbol{\varepsilon}$, and the independent arm draws row v as $\sigma_v \boldsymbol{\varepsilon}_v$ with $\sigma_v^2 = \sum_u T_{vu}^2$ taken from the implemented factor, so the row laws agree to floating-point precision even where the cached eigenbasis is not exactly orthonormal. Each of ten trials shares the split and the trainable initialisation, which a hash verifies. The endpoints, recorded in the driver before any run, are base-mode accuracy, thirty-sample node NLL, and entropy AUROC under matched graph-level noise, with margins of 0.005, 0.01 nats, and 0.01, and Holm correction applies separately to the nine difference tests and the nine equivalence tests. Panels (b) to (d) of Figure 8 give the nine contrasts. Accuracy on WikiCS and Amazon Computers, node NLL on Amazon Computers, and AUROC on Cora pass the equivalence test after correction. The other five cells are inconclusive, since every interval contains zero but the wider ones still cross the margins at ten trials.

F.3 Joint Structure of the Predictive Law

Shape cannot affect per-node laws at fixed weights, but it can affect the joint law. Figure 9 correlates the per-node top-class-probability fluctuations across 500 joint draws of the trained full-spectrum model on Cora. Neighbours co-fluctuate well above the Monte Carlo floor, and the correlation decays to that floor within two hops, whereas the variance-matched isotropic prior stays at the floor at every distance (panel a). The BFS-ordered correlation matrix shows adjacency-aligned blocks only under the GMRF (panels b and c). The fluctuation energy concentrates on the smooth Laplacian modes and falls below the isotropic reference at high frequencies (panel d), which is the spectrum the prior imposes, so its signature persists through the conditioned flow. The fluctuation itself is very small. A separate 200-draw decomposition of predictive label counts on 50-node Cora patches, which treats labels as conditionally independent given the draw, finds that off-diagonal covariance is a sizeable

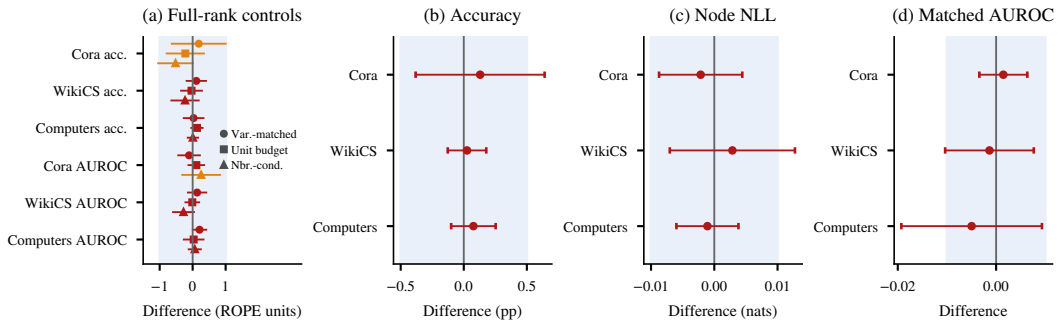

Figure 8: Full-rank controls. (a) Paired 90% intervals of the variance-matched, unit-budget, and neighbour-conditioned contrasts in ROPE units, orange where the Holm-corrected equivalence test fails. (b–d) Exact marginal matching, structured minus independent rows over ten paired trials, with unadjusted 95% intervals shaded over the margins.

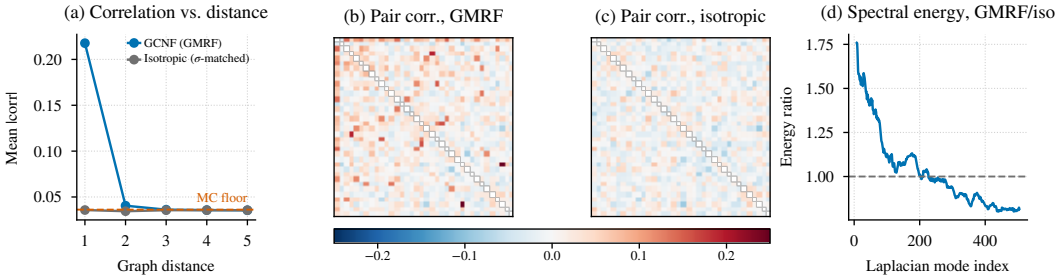

Figure 9: Normalised dependence in the joint predictive law on Cora over 500 draws. (a) Correlation against graph distance. (b, c) Pairwise correlation, BFS-ordered. (d) Energy per Laplacian mode.

share of the latent-sampling component but a negligible share of the total variance, because the trained network absorbs the draw (Appendix E.3).

G Crossing the Objective Boundary

Every arm of Appendix F keeps the separable objective of Eq. (9), so its null results cannot distinguish a useless prior from one that the objective never rewards. This appendix removes the separable premise, crosses it with the node-wise premise, and measures how much cross-node dependence either route can reach in principle.

G.1 A Pairwise Objective and a Latent-Mixing Decoder

We retrain the four-coupling model with the pairwise composite log score (Varin et al., 2011; Gneiting & Raftery, 2007)

$$\mathcal{L}_{\text{pair}} = -\frac{1}{|\mathcal{E}_{\text{tr}}|} \sum_{(u,v)\in\mathcal{E}_{\text{tr}}} \log\Big[\frac{1}{S}\sum_{s=1}^{S} p_s(y_u \mid u)\, p_s(y_v \mid v)\Big] \tag{44}$$

over training edges with both endpoints labelled, where $p_s(y_v \mid v)$ is the probability of the observed label at v under draw s and both endpoints share the draw index, at $S = 4$ and unit weight. Sharing s is the only channel through which the correlation between rows u and v of the prior can enter a loss, and Proposition 6.1 does not bind Eq. (44) because the term is not separable. The dependence that the predictive law realises is

$$\text{dep} = \frac{1}{|\mathcal{E}_{\text{te}}|} \sum_{(u,v)\in\mathcal{E}_{\text{te}}} \log \frac{\frac{1}{S}\sum_s p_s(y_u \mid u)\, p_s(y_v \mid v)}{\big(\frac{1}{S}\sum_s p_s(y_u \mid u)\big)\big(\frac{1}{S}\sum_s p_s(y_v \mid v)\big)}, \tag{45}$$

the gap in nats per test edge between the sampled joint predictive and the product of its marginals at $S = 30$, which no per-node endpoint sees. The draw average of $p_s(y_u \mid u)\, p_s(y_v \mid v)$ equals the

product m_{uv} of the two draw averages plus their covariance across draws, so Eq. (44) falls as that covariance grows, and each term of Eq. (45) equals $\log(1 + \mathrm{cov}_{uv}/m_{uv})$. Its population counterpart vanishes when the predictive laws at u and v are independent, which an isotropic prior forces under a node-wise decoder, but the plug-in value is biased at finite S, so we compare the structured arm with the isotropic arm rather than with zero.

Factorial Design. We cross prior, objective, and decoder at two levels each on Cora, WikiCS, and Amazon Computers over ten trials. The prior factor sets the full-rank GMRF against isotropic noise rescaled to the mean marginal variance the structured arm learns in the same cell, so the two share the budget of Proposition C.3 and differ in profile and shape. The objective factor adds Eq. (44) or not. Every cell shares one marginal term, the S-draw mean-probability negative log-likelihood plus the auxiliary probe, so the separable cells are matched controls for the joint cells rather than reruns of Section 6.4. The decoder factor sets the node-wise coupling against one whose condition also reads the mean of the neighbours' untouched half, which makes the predictive law at v depend on its neighbours' latent rows. One routine trains all eight cells, so optimiser, schedule, validation rule, draw counts, and evaluation are identical by construction.

Manipulation Check. The coupling's scale and translation heads start at zero and no separable term rewards the neighbour mean, so the mixing decoder could learn to ignore it. We therefore replace the neighbour term inside each condition by the same term computed from an independent draw, which keeps its law and destroys its alignment with the graph. The substitution changes the per-node predictive, which it would leave exactly unchanged under a node-wise decoder, so the channel is in use. Resampling the whole latent changes the same predictive at least an order of magnitude more, and even that change is small in total variation, so the neighbour route carries a small share of a stochastic channel that is itself nearly inactive, as Eq. (39) predicts. The pairwise objective increases both quantities.

Primary Endpoints. We fixed two endpoints before the runs, both in held-out pairwise NLL and each required to reach the 5% level with the structured arm favoured on at least two of the three graphs. The objective-by-prior interaction among the node-wise cells passes on no graph, and all three point estimates carry the sign a useless prior would produce. The structured-minus-isotropic contrast inside the cell that combines the pairwise objective with the mixing decoder separates on no graph either, with two of three point estimates of that sign (Figure 10a), and the three-way interaction that "both routes are required" predicts is indistinguishable from zero on every graph. None of the forty-eight secondary comparisons of accuracy, node NLL, AUROC-H, and AURC-H across graphs and cells reaches the 5% level (panel d). Against the regions of Appendix F.1, with ± 0.01 nats declared for the pairwise likelihood, all nine contrast-metric cells that miss equivalence after Holm correction are pairwise-likelihood cells whose wide intervals reflect imprecision rather than an effect.

Realised Dependence. The design changes the realised dependence on every graph (Figure 10b). With the node-wise decoder the pairwise objective increases it severalfold, and on Cora the structured prior then realises significantly more than its control. The mixing decoder alone changes it less and not always upwards, and the cell that crosses both premises is the largest on every graph. Even that value stays more than thirty times below the 0.01-nat margin of the pairwise endpoint. A covariance across draws cannot exceed the product of the spreads of the two probabilities it couples, and each node's prediction varies little between draws, so every incentive we supply widens a channel that remains too narrow to change any endpoint.

G.2 How Much Dependence Is Reachable at All

A task whose labels carry no cross-node dependence cannot reward a prior that supplies one, so we build a task in which dependence is the only learnable signal and score prescribed decoders on it exactly, with no training and no sampling.

Task Construction. Each of ten generated stochastic block graphs has $n = 2000$ nodes in 40 blocks of 50, with edge probability 0.10 within blocks and 5×10^{-4} between them. Node v carries an observed bit b_v beside 32 noisy prototype features, and its label is $y_v = b_v \oplus g_{\mathrm{block}(v)}$ for a fair block coin g redrawn on every use. Every label is uniform given the graph and the features, which fixes every per-node endpoint. Within a block, however, $y_u \oplus y_v = b_u \oplus b_v$, so the optimal pairwise NLL is $\log 2$ per within-block edge against $\log 4$ for independent predictions, which leaves a channel of 0.693 nats that only cross-node dependence can reach.

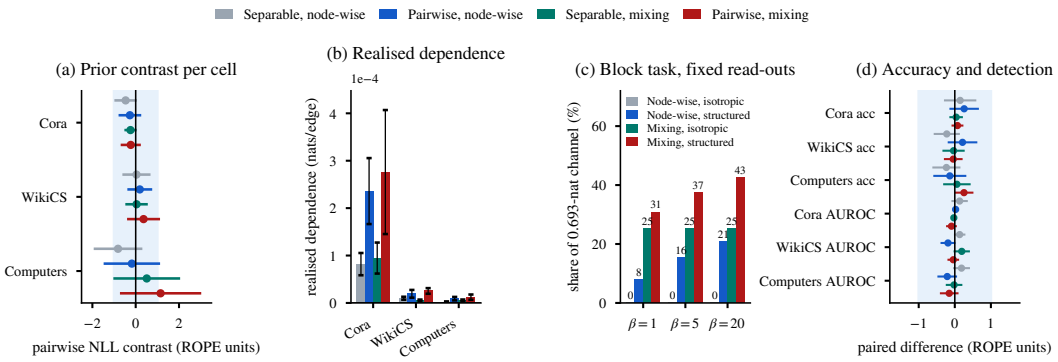

Figure 10: The prior × objective × decoder factorial. (a) Structured minus isotropic pairwise NLL per cell, as 90% intervals in ROPE units. (b) Realised dependence per cell. (c) Channel shares of the fixed sign read-out on the block task. (d) Accuracy and detection contrasts in ROPE units.

Table 11: Exact pairwise NLL of the fixed sign read-outs, in nats per within-block edge, as mean±standard error across ten generated block graphs. Averaging fixes $\gamma = 1$, and the node-wise independent score is exactly $\log 4$.

β	Mean marginal variance	Node-wise read-out		Neighbour averaging	
		Independent	Structured	Independent	Structured
1	0.524	$\log 4$	1.3310 ± 0.0002	1.2105 ± 0.0005	1.1724 ± 0.0006
5	0.196	$\log 4$	1.2784 ± 0.0004	1.2105 ± 0.0005	1.1274 ± 0.0007
20	0.062	$\log 4$	1.2421 ± 0.0006	1.2105 ± 0.0005	1.0905 ± 0.0009

Fixed Read-Outs. Let z be one coordinate of a full-spectrum Gaussian prior with covariance $\Sigma = (I + \beta\tilde{\boldsymbol{L}})^{-1}$, and standardise $\tilde{z}_v = z_v/\sigma_v$ with $\sigma_v = \Sigma_{vv}^{1/2}$. The node-wise read-out

$$p(y_v = b_v \mid z) = \mathbf{1}\{\tilde{z}_v > 0\} \tag{46}$$

reads the coordinate at v alone, and its independent control draws $z'_v = \sigma_v \varepsilon_v$ with the same marginals. With $R = D^{-1}A$, whose row is zero at an isolated node, the averaging read-out

$$s = (I + \gamma R)\,\tilde{z}, \qquad p(y_v = b_v \mid z) = \mathbf{1}\{s_v > 0\} \tag{47}$$

returns Eq. (46) at $\gamma = 0$. The primary comparison fixes $\gamma = 1$ for every graph and damping, and no temperature, normalisation constant, or parameter is estimated from sampled outputs.

Exact Population Score. Let C be the correlation matrix of the prior, or I for the control, so that the aggregated covariance is

$$V_\gamma = (I + \gamma R)\, C\, (I + \gamma R)^\top. \tag{48}$$

A centred Gaussian pair with correlation r_{uv} puts probability $\frac{1}{4} + \arcsin(r_{uv})/(2\pi)$ on each of the two configurations the task allows, so the expected pairwise NLL on a within-block edge is exactly

$$-\log\Big[\frac{1}{4} + \frac{\arcsin(r_{uv})}{2\pi}\Big]. \tag{49}$$

Averaging over the within-block edges gives the score of each graph with no Monte Carlo error (Table 11).

Reachable Share. A node-wise read-out of the structured prior exceeds the $\log 4$ floor of every independent node-wise arm far outside the spread across graphs, so the pairwise endpoint has power and the prior carries cross-node information that a per-node read-out can reach once the score stops being separable. That share remains a minority of the channel, even on graphs whose blocks are what the smooth modes of the prior describe, and it grows slowly with the damping (Figure 10c). The score reads the prior only through the correlation r_{uv} of the two standardised draws on an edge, and a GMRF raises that correlation only by damping its rough modes, so a larger share requires a prior that removes most of its own variance (Table 11).

Comparison of the Two Routes. Averaging over neighbours creates dependence from independent rows, because the two ends of a within-block edge share terms, and at $\gamma = 1$ it reaches a larger share

Table 12: Training efficiency of GCNF as parameter count and per-epoch training time, where time is the steady-state mean of each step. The eigenpair cost is one-time preprocessing and is excluded from training time.

Dataset	Nodes	Params	Train (s/ep)	Eig. (s)
Cora	2,708	1.48M	0.028	18
CiteSeer	3,327	1.84M	0.029	22
PubMed	19,717	1.33M	0.045	547
Amazon Photo	7,650	1.37M	0.045	62
Amazon Computers	13,752	1.37M	0.074	239
Coauthor CS	18,333	2.34M	0.043	423
Coauthor Physics	34,493	2.60M	0.088	58
WikiCS	11,701	1.30M	0.086	137

than the prior reaches alone at any damping we sweep. Crossing the two routes reaches more than either alone and less than their sum, since the averaging read-out captures part of the dependence the prior would otherwise supply. At $\beta = 5$ the prior's value falls from 0.108 to 0.083 nats once the decoder averages at $\gamma = 1$, and both orderings hold at every $\gamma \in \{0.25, 0.5, 1, 2, 4\}$. Over the range we can evaluate, the two routes are therefore substitutes, and the cheaper one needs no spectral decomposition and no damping to tune. These scores describe prescribed decoders, and benchmark training need not attain them.

H Implementation and Efficiency

Hardware and Software. We train on one NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory. The software stack is Python 3.11, PyTorch 2.11, PyTorch Geometric 2.7, scikit-learn 1.8, and SciPy 1.17.

Architecture Details. Both input maps, the context map ϕ_h and the flow input ϕ_z of Eq. (8), are per-node linear layers followed by LayerNorm. The spectral positional encoding is a two-layer MLP on the k smoothest eigenvectors U_k, added to $H^{(0)}$ through the learnable gate $\alpha = e^{\gamma}$. DCE uses GATConv with concatenated heads and attention dropout 0.1, and it applies LayerNorm after the residual sum. The coupling networks MLP_s and MLP_t use GELU activations and a linear output layer initialised to zero weights and biases, so every coupling starts at the identity, and the classifier ψ has one hidden GELU layer.

Eigenpair Computation. Under our convention that $D_{vv}^{-1/2} = 0$ at isolated nodes, those nodes have Laplacian eigenvalue one, so the number of zero eigenvalues is the number of connected components that contain an edge, where a self-loop, which WikiCS keeps at 520 nodes, counts as an edge. Cora has 78 such components, CiteSeer 390, WikiCS 22, Amazon Photo 21, and Amazon Computers 33, and two of these counts exceed k, so the solver counts the null space through the connected components and returns the k smallest eigenpairs *above* it. Graphs up to $20{,}000$ nodes use dense `eigh`, Coauthor Physics uses a dense GPU float32 solve with a float64 CPU fallback, and ogbn-arxiv runs GPU LOBPCG on its giant component and dense solves on the small ones. Every basis passes the relative residual check $\|\tilde{\boldsymbol{L}}v - \lambda v\|/(1+\lambda) \le 5 \times 10^{-3}$ and is cached once per graph, shared across trials and variants, and excluded from training times.

H.1 Efficiency

A forward pass of GCNF costs

$$O\Big(nf(d+d_h) + ndk + nk_{\mathrm{LPE}}d_h + nd_h^2 + K\big(|\mathcal{E}|\,d_h + n\,d_h^2 + n\,\nu\,(d+d_h+\nu)\big)\Big), \tag{50}$$

with ν the coupling-MLP width and $k_{\mathrm{LPE}} \le k$ the rank of the positional encoding. The input projections contribute $O(nf(d+d_h))$, the positional encoding $O(nk_{\mathrm{LPE}}d_h + nd_h^2)$, the spectral projection $O(ndk)$, each attention layer $O(nd_h^2 + |\mathcal{E}|d_h)$, and each coupling $O(n\nu(d+d_h+\nu))$, so the pass is linear in $|\mathcal{E}|$. Given the context, the three density terms of Eq. (43) cost $O(ndk + nd^2K)$ at widths $\nu, d_h = O(d)$.

Training Cost. Table 12 gives GCNF's parameters, its training time per epoch, and the one-time eigenpair cost, which covers component analysis, factorisation, and residual validation on a shared

machine and is indicative rather than a scaling curve. The eigenpair cost is not monotone in n, because Coauthor Physics crosses onto the dense GPU solver, which finishes much sooner than the CPU factorisation of PubMed. Input feature dimension, rather than graph size, governs the parameter count.

Efficiency of DCE Versus GCNF. We time optimiser steps on cached spectra without validation, as the median of five windows of ten steps after ten warm-up steps. GCNF carries 2.2 to 7.8 times DCE's parameters, and the ratio is smallest on the high-dimensional co-authorship graphs, where the input projections that both models carry dominate the count. A DCE training step still takes more than half of GCNF's time, which is consistent with a step that the shared attention layers dominate on the denser graphs. At inference DCE scores a node in one forward pass, whereas GCNF's sampled uncertainty repeats the pass for each of its thirty draws.