INSTANCE INDISTINGUISHABILITY AND THE LIMITS OF GRAPH LEARNING: A LOCAL-GLOBAL PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph-level prediction infers global properties from local neighborhoods, but it may fail when classes differ globally while local observations remain indistinguishable. We ask when interactions among local observations reveal a global difference and whether graph learners can exploit it. Planted Model RB makes this question controllable by varying problem size and constraint density. Although satisfiability distinguishes planted from random instances above the classical threshold, every individual constraint has the same distribution in both models. To characterize this transition, we derive the low-degree likelihood-ratio projection, showing that the first informative component comes from constraint pairs sharing variables and giving its exact finite-sample strength. We prove further that, when this pairwise signal vanishes asymptotically, all polynomial statistics up to logarithmic degree are uninformative, yielding a local-global separation that widens with problem size. Experiments reveal a corresponding but steeper size dependence in graph learning. Over the sampled range, a regime generally widens in which shared-variable evidence is detectable from the complete instance but is not effectively extracted by the evaluated GNNs. This discrepancy is reduced more consistently by representations that preserve shared-variable relations than by generic structural augmentation or the evaluated increases in architectural expressivity. Our theoretical and empirical findings therefore indicate that learning a global distinction from local observations depends on both sufficient overlap among those observations and a representation that preserves the structure through which the distinguishing signal emerges.

1 INTRODUCTION

Graph-level prediction assigns a label to an entire graph, yet most scalable graph neural networks (GNNs) construct that prediction by repeatedly aggregating local neighborhoods (Xu et al., 2019; Morris et al., 2019). This local-to-global design is naturally suited to problems in which the class distinction recurs across the neighborhoods being aggregated. It becomes less reliable when the label depends on a global relation that is absent from every individual local observation. For example, a connected six-cycle and two disjoint triangles differ in global connectivity, although uniform node features give every vertex the same local multiset, as illustrated in Figure 1. The one-dimensional Weisfeiler–Leman test (1-WL) consequently assigns identical color histograms to the two graphs at every iteration. Standard message passing therefore produces the same graph representations and cannot recover the global distinction from these local summaries.

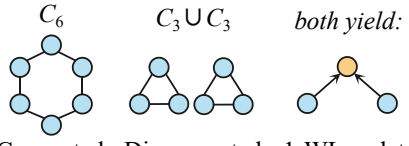


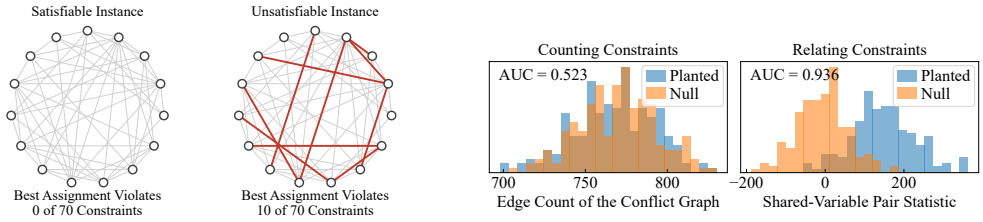
Figure 1: Graphs with different connectivity that 1-WL cannot distinguish.

To study this form of instance indistinguishability from a local-global perspective, we use planted Model RB to construct a family in which problem size and constraint density are controllable. Model RB is a random constraint-satisfaction model with a polynomially growing domain and an exact asymptotic satisfiability threshold (Xu & Li, 2000). We compare null instances with forced-satisfiable instances generated from a hidden assignment (Xu et al., 2007). Above the threshold, a null instance is unsatisfiable with probability tending to one, whereas a planted instance is satisfiable by construction. Once the hidden assignment is marginalized, however, every individual constraint has exactly the

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.

same distribution in the two classes. The label is therefore global, while single-constraint evidence is absent by construction.

The case study in Figure 2 shows the constraint-scope graphs at $n = 15$, where nodes represent variables, gray edges represent constraints, and red edges mark constraints violated by the displayed assignment. The planted instance admits a satisfying assignment, whereas exact search certifies that the null instance is unsatisfiable. Nevertheless, the coarse *Counting Constraints* summary attains only 0.523 AUC across 200 instances per class. The *Relating Constraints* statistic supplements these local observations with their shared-variable relations and reaches 0.936 AUC. This contrast illustrates the central difficulty: local observations need not represent a global label, and recovering the distinction may require modeling how those observations interact.



(a) Scope graphs of two sampled instances (b) Distributions of two instance statistics

Figure 2: Model RB case study at $n = 15$. (a) Scope graphs of a satisfiable planted instance and an unsatisfiable null instance. (b) Distributions of the two statistics over 200 instances per class.

This construction raises three questions that are often conflated: (i) when are the two instance distributions globally distinguishable, which concerns *satisfiability*; (ii) at what density does the distinction become visible to bounded-order statistics, which concerns *statistical visibility*; and (iii) once such evidence exists, can a graph representation and learning algorithm preserve and extract it, which concerns *empirical learnability*. Separating these questions distinguishes the absence of bounded-order signal from the failure of a learning pipeline to use signal that is already present.

In this paper, we first characterize when the global distinction becomes visible to bounded-order statistics by combining the classical satisfiability threshold with a low-degree analysis. The degree-one signal vanishes exactly, whereas the first informative component arises from constraint pairs sharing variables. We derive its exact finite-sample strength Λ_2, identify it as the optimal squared SNR among quadratic detectors, and prove a local-global separation through logarithmic degree that widens with problem size. This result concerns the complete constraint instance rather than a GNN-specific lower bound. We then examine whether graph learners can exploit the available evidence. A shared-variable detector follows the scale predicted by Λ_2, whereas the evaluated GNNs exhibit a steeper size-dependent learning crossover. The resulting extraction gap narrows more consistently when the representation preserves shared-variable relations than under the generic structural and architectural changes evaluated in this work. The contributions of this work are as follows.

- We use planted Model RB to formulate a controlled local-global view of instance indistinguishability, with a global satisfiability distinction but identical single-constraint marginals.
- We derive the exact low-degree signal Λ_2 and prove a size-dependent separation between global distinguishability and polynomial statistics up to logarithmic degree.
- We identify a widening empirical extraction gap and show that preserving shared-variable relations improves learnability more consistently than the generic representation and architecture changes evaluated in this work.

2 Problem Setup and Preliminaries

2.1 Finite Model RB

We use $[u] = \{1, \ldots, u\}$ and write $\binom{[n]}{k}$ for the collection of k-subsets of $[n]$. A constraint has scope $S_a \in \binom{[n]}{k}$ and allowed relation $R_a \subseteq [q]^k$, and the complete instance is $X = (S, R) = ((S_a, R_a))_{a=1}^m$.

Definition 2.1 (Finite Model RB parameters)**.** Fix integers $n \geq k \geq 2$, $q \geq 2$, $m \geq 0$, and $1 \leq s < q^k$. Let

$$N := q^k, \qquad \rho := \frac{s}{N}, \qquad p_{\text{eff}} := 1 - \rho, \qquad \omega := \frac{p_{\text{eff}}}{\rho}. \tag{1}$$

Each relation table admits exactly s tuples and forbids exactly $N - s$ tuples.

The original notation uses $d = n^\alpha$, $m = rn \log n$, and pd^k forbidden tuples per constraint (Xu & Li, 2000). Here q is the realized integer domain size and $p_{\text{eff}} = (N - s)/N$ is the realized tightness. Exact finite-sample formulas use q, m, s, and p_{eff}, whereas asymptotic statements assume $q_n \sim n^\alpha$, $m_n \sim rn \log n$, and $p_{\text{eff},n} \to p \in (0, 1)$ under any fixed rounding convention. We write $\omega = p_{\text{eff}}/(1 - p_{\text{eff}})$.

Here n and m count CSP variables and constraints, whereas the graph has $|\mathcal{V}| = nq$ value nodes and a derived edge set. Thus $n \neq |\mathcal{V}|$ and $m \neq |\mathcal{E}|$. Appendix A gives complete notation.

Assumption 1 (Random scopes)**.** The scopes $S_1, \ldots, S_m$ are independent and uniform on $\binom{[n]}{k}$. Variables within a scope are distinct, while repeated scopes across constraints are allowed. The realized scopes are observed under both hypotheses.

2.2 Null and planted distributions

Definition 2.2 (Null distribution)**.** Under $\mathbb{P}_0$, the scopes are first sampled according to Assumption 1. Conditional on S, the relations $R_1, \ldots, R_m$ are sampled independently, with each R_a uniform over all size-s subsets of $[q]^k$.

Definition 2.3 (Mixed planted distribution)**.** Under $\mathbb{P}_1$, sample $\sigma^* \sim \text{Unif}([q]^n)$ and the scopes from Assumption 1. Conditional on (σ^*, S), the relations are independent and uniform among size-s relations satisfying $\sigma^*_{S_a} \in R_a$. The hidden assignment is then marginalized, so $\mathbb{P}_1$ is a mixture rather than a model that reveals σ^*.

The two distributions share the scope law, the number of constraints, and the relation-table size. Given one observation, the detection task determines whether it was drawn from $\mathbb{P}_0$ or $\mathbb{P}_1$ without requiring an estimator to recover σ^*.

Let $Z(X)$ denote the number of satisfying assignments of X. The planted class satisfies $Z(X) \geq 1$ almost surely. The following classical result identifies when satisfiability separates the two distributions globally.

Proposition 2.4 (Classical Model RB threshold Xu & Li, 2000)**.** Suppose $q = n^\alpha$, $m = rn \log n$, $\alpha > 1/k$, and $k \geq 1/(1 - p)$, with fixed k, α, and $p \in (0, 1)$. The null model has the asymptotic satisfiability threshold

$$r_{\text{cr}} = \frac{\alpha}{-\log(1 - p)}. \tag{2}$$

For fixed $r > r_{\text{cr}}$, a null instance is unsatisfiable with probability tending to one. For fixed $r < r_{\text{cr}}$, it is satisfiable with probability tending to one.

For $k = 2$, the condition $k \geq 1/(1 - p)$ becomes $p \leq 1/2$. Eq. (2) is asymptotic and therefore differs from a finite-n crossover. Parameter settings outside these conditions are not covered by the exact-threshold theorem.

2.3 Observation, graph encoding, and low-degree projection

Let $L_X(X) := \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}(X)$ be the likelihood ratio of the complete observation. A graph representation is a deterministic map $G = (\mathcal{V}, \mathcal{E}) = \phi(X)$, such as a relation-aware factor graph, a multigraph, or a deduplicated microstructure graph. The induced graph distributions are $\mathbb{P}_i^G = \phi_\# \mathbb{P}_i$ for $i \in \{0, 1\}$.

Proposition 2.5 (Likelihood ratio after deterministic encoding)**.** Assume $\mathbb{P}_1 \ll \mathbb{P}_0$. The graph-level likelihood ratio satisfies

$$L_G(G) := \frac{\mathrm{d}\mathbb{P}_1^G}{\mathrm{d}\mathbb{P}_0^G}(G) = \mathbb{E}_0[L_X(X) \mid G]. \tag{3}$$

Consequently,

$$\chi^2(\mathbb{P}_1^G \| \mathbb{P}_0^G) \leq \chi^2(\mathbb{P}_1 \| \mathbb{P}_0). \tag{4}$$

Appendix C.1 restates the proposition and gives its proof.

Eq. (4) concerns total chi-square information. A nonlinear encoding need not preserve polynomial degree, so it does not compare the raw degree-two norm directly with the degree-two norm of a graph representation.

For every constraint a and tuple $z \in [q]^k$, define $Y_{a,z} := \mathbf{1}\{z \in R_a\}$. Conditional on a fixed scope configuration $S = s_0$, let $\mathcal{H}_{\le D}(s_0)$ be the subspace of $L^2(\mathbb{P}_0^{R|S=s_0})$ spanned by polynomials of total degree at most D in the coordinates $\{Y_{a,z}\}$. Because the scopes are observed side information, they do not count toward degree. We define the joint low-degree projection fiberwise:

$$L_{\le D}(s_0, \cdot) := \mathrm{Proj}_{\mathcal{H}_{\le D}(s_0)} L_X(s_0, \cdot), \tag{5}$$

so that

$$\|L_{\le D}\|_2^2 = \mathbb{E}_S\left[\|L_{\le D}(S, \cdot)\|^2_{L^2(\mathbb{P}_0^{R|S})}\right]. \tag{6}$$

We write $\mathcal{H}_{\le D}$ for the corresponding joint fiberwise class. A measurable statistic $f(S, R)$ belongs to $\mathcal{H}_{\le D}$ when $f(s_0, \cdot) \in \mathcal{H}_{\le D}(s_0)$ for almost every scope configuration s_0. The coefficients of a polynomial may therefore depend measurably on the observed scopes, while its degree counts only relation-table indicators. All exact low-degree statements in Section 3 concern this joint random-scope observation. They are not pointwise claims for every fixed scope multiset.

3 Low-Degree Detectability of the Global Witness

We characterize when the planted assignment becomes visible in the complete random-scope observation. Appendix C restates all results and gives the proofs. Throughout, $N = q^k$, $\rho = s/N$, $p_{\mathrm{eff}} = 1 - \rho$, and $\omega = p_{\mathrm{eff}}/\rho$.

3.1 Exact quiet planting

The unrestricted likelihood ratio has a direct combinatorial form.

Theorem 3.1 (Likelihood ratio equals normalized solution count)**.** For the mixed planted and null distributions,

$$L_X(X) = \frac{Z(X)}{q^n \rho^m} = \frac{Z(X)}{\mathbb{E}_0 Z(X)}. \tag{7}$$

Let $H \sim \mathrm{Bin}(n, 1/q)$ be the agreement count of two independent assignments and set $\gamma_H := \omega\{q^k (H)_k/(n)_k - 1\}/(q^k - 1)$.

Theorem 3.2 (Exact low-degree norm)**.** For every integer $D \ge 0$,

$$\|L_{\le D}\|_2^2 = \mathbb{E}_H\left[\sum_{t=0}^{\min\{D,m\}} \binom{m}{t} \gamma_H^t\right]. \tag{8}$$

The degree-one term in Eq. (8) vanishes exactly.

Corollary 3.3 (Exact quiet planting)**.** At every finite n, $(L_X)_{=1} = 0$ and $\|L_{\le 1}\|_2^2 = 1$. Equivalently, the mixed planted and null distributions have identical one-constraint marginals.

Thus the global witness leaves no unary trace after the planted assignment is marginalized. Detectability begins only through consistency across multiple constraints.

3.2 The first signal is pairwise

Let J be the overlap size of two independent uniform scopes. Its law is

$$\Pr(J = j) = \frac{\binom{k}{j}\binom{n-k}{k-j}}{\binom{n}{k}}, \qquad j = 0, \dots, k. \tag{9}$$

Define the average contribution of one constraint pair and the accumulated second-order signal by

$$B_{n,q,k,p_{\mathrm{eff}}} := \frac{\omega^2}{(q^k-1)^2} \sum_{j=1}^{k} \Pr(J=j)(q^j-1), \tag{10}$$

$$\Lambda_2 := \binom{m}{2} B_{n,q,k,p_{\mathrm{eff}}}. \tag{11}$$

Theorem 3.4 (Exact pairwise signal and optimal SNR)**.** The excess degree-two likelihood-ratio norm is

$$\|L_{\le 2}\|_2^2 - 1 = \Lambda_2. \tag{12}$$

Moreover,

$$\sup_{\substack{f \in \mathcal{H}_{\le 2},\ \mathbb{E}_0 f = 0 \\ \mathrm{Var}_0(f) > 0}} \frac{(\mathbb{E}_1 f - \mathbb{E}_0 f)^2}{\mathrm{Var}_0(f)} = \Lambda_2. \tag{13}$$

Hence $\sqrt{\Lambda_2}$ is the largest null-standardized mean separation attainable by a quadratic detector on the complete observation.

Degree here counts centered relation-table indicators, not GNN depth or WL order. The constructive optimizer sums centered consistency over overlapping constraint pairs (Appendix C). For $k = 2$, its dominant term groups relation marginals by shared variable, while the exact score also handles repeated scopes. This yields a polynomial-time diagnostic detector, not a satisfiability solver or planted-assignment estimator.

The factor $\binom{m}{2}$ counts opportunities for constraints to interact, while $B_{n,q,k,p_{\mathrm{eff}}}$ measures the mean squared signal per pair. For any fixed $\eta > 0$, the contour $\Lambda_2 = \eta$ is therefore an exact constant-SNR boundary, with real-valued constraint count

$$m_{2,\eta} = \frac{1 + \sqrt{1 + 8\eta / B_{n,q,k,p_{\mathrm{eff}}}}}{2}. \tag{14}$$

The reference $\eta = 1$ is a unit-SNR convention rather than a parameter of Model RB.

Corollary 3.5 (Finite arity-two contour)**.** For $k = 2$ and independent scopes sampled with replacement,

$$B_{n,q,2,p_{\mathrm{eff}}} = 2 \left(\frac{p_{\mathrm{eff}}}{1 - p_{\mathrm{eff}}} \right)^2 \frac{2n + q - 3}{n(n-1)(q-1)(q+1)^2}. \tag{15}$$

Together with Eq. (14), this gives an exact finite-instance density contour for every $(n, q, p_{\mathrm{eff}}, \eta)$. The explicit unit-SNR form is given in Appendix C.9.

3.3 Visibility scale and the satisfiability threshold

The finite contour yields a simple asymptotic scale.

Proposition 3.6 (Asymptotic low-degree scale)**.** Fix $k \ge 2$, suppose $q = o(n)$, $p_{\mathrm{eff},n} \to p \in (0, 1)$, and let $m \to \infty$. Then

$$\Lambda_2 = \frac{k^2}{2} \left(\frac{p}{1-p} \right)^2 \frac{m^2}{n q^{2k-1}} (1 + o(1)). \tag{16}$$

If $q = n^\alpha$ and $m = rn \log n$, the unit-SNR density satisfies

$$r_{\mathrm{LD}}(n) \sim \frac{\sqrt{2}(1-p)}{kp} \frac{n^{\alpha(k-1/2)-1/2}}{\log n}. \tag{17}$$

Eq. (16) identifies the source of the signal: $\Lambda_2 \asymp (m^2/n) q^{-(2k-1)}$, combining the number of overlapping constraint pairs with the evidence per pair. Their degree-two blocks are orthogonal under the null, although the pairs need not be independent. Increasing r therefore adds repeated consistency checks without changing one-constraint marginals, while increasing α weakens each check through the larger domain. We interpret Λ_2 as accumulated *redundant consistency* in the complete instance, without assuming that an encoding or learner preserves it.

Corollary 3.7 (Arity-two separation from satisfiability)**.** Suppose $k = 2$, $1/2 < \alpha < 1$, $p \le 1/2$, and the classical Model RB threshold assumptions hold. Then

$$\frac{r_{\mathrm{LD}}(n)}{r_{\mathrm{cr}}} \sim \frac{(1-p)[-\log(1-p)]}{\sqrt{2}\,p\alpha} \frac{n^{(3\alpha-1)/2}}{\log n} \longrightarrow \infty. \tag{18}$$

The two densities measure different events. Above r_{cr}, satisfiability separates the distributions by a global witness. Around $r_{\mathrm{LD}}(n)$, repeated overlaps accumulate constant low-order SNR. Their diverging ratio quantifies a widening interval in which the label is globally decisive but remains invisible to bounded-order statistics.

3.4 Logarithmic-degree invisibility

The pairwise scale also controls every polynomial degree up to logarithmic order.

Theorem 3.8 (Logarithmic-degree invisibility)**.** Fix $k \ge 2$ and $1/k < \alpha < 1$. Let $q = q_n$ be an integer sequence satisfying $q_n \sim n^\alpha$, assume $p_{\mathrm{eff},n} \to p \in (0,1)$ while staying away from 0 and 1, and let $m = m(n)$ be an integer sequence for which $m(n) \le n^A$ eventually for some fixed $A < \infty$. For every fixed $C_D > 0$ and every integer sequence $D = D(n) \le C_D \log n$,

$$\Lambda_2 \to 0 \quad \Longrightarrow \quad \|L_{\le D}\|_2^2 = 1 + o(1). \tag{19}$$

Conversely, if $D \ge 2$ and $\Lambda_2 \to \infty$, then

$$\|L_{\le D}\|_2^2 \to \infty. \tag{20}$$

Corollary 3.9 (Global distinguishability before low-degree visibility)**.** Assume the classical threshold conditions, $1/k < \alpha < 1$, $q_n \sim n^\alpha$, $m_n \sim rn \log n$ for fixed $r > r_{\mathrm{cr}}$, and relation sizes chosen so that $p_{\mathrm{eff},n} \to p \in (0,1)$. Satisfiability distinguishes $\mathbb{P}_1$ from $\mathbb{P}_0$ with error tending to zero, whereas for every fixed $C_D > 0$,

$$\|L_{\le C_D \log n}\|_2^2 = 1 + o(1). \tag{21}$$

Consequently, every centered degree-$C_D \log n$ statistic with unit null variance has vanishing planted-versus-null mean separation.

Corollary 3.9 formalizes the central local-global separation. Satisfiability provides a global semantic distinction, quiet planting eliminates unary evidence, and overlapping constraints gradually accumulate low-degree consistency. These are properties of the complete instance distributions, whereas a trained GNN must additionally preserve and extract the evidence through its graph encoding. Extending the result to general polynomial-time algorithms requires the low-degree conjecture (Kunisky et al., 2019), and relating it to a particular GNN requires separate analysis of encoding loss, WL order, communication, optimization, and data. Thus, even when complete instances are globally distinguishable, local information may fail to represent their global semantics.

4 Experiments

We examine how problem size, statistical evidence, and graph representation govern the learnability of the global distinction. We ask:

- **RQ1:** How does the graph-learning crossover vary with problem size and constraint density?
- **RQ2:** Does Λ_2 predict the finite-sample visibility of pairwise evidence?
- **RQ3:** What limits signal extraction, and which interventions recover it?

Experimental Settings and Protocols. We use binary constraints ($k = 2$) and sweep $\alpha \in \{0.6, 0.8\}$, $p \in \{0.5, 0.3\}$, $n \in \{15, 25, 40, 60\}$, and $r \in \{0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 6\} \cdot r_{\mathrm{cr}}(\alpha, p)$. Finite quantities use realized q, m, and p_{eff}. Within each balanced cell, null and planted instances share (n, q, m, t) and the scope law, and the planted assignment is never exposed. We use the deduplicated microstructure graph described in Appendix D.2, with one node per variable–value pair and normalized degree as its initial feature; constraint identities, scope labels, and edge multiplicities are not provided. The main classifier is a 3-layer GIN (Xu et al., 2019) with hidden width 64 and dropout 0.5.

Table 1: GIN test accuracy and AUC on the complete main grid, in percent (mean ± standard deviation over three seeds).

(α, p)	n	Metric	r/r_{cr} 0.5	0.75	1	1.5	2	3	4	6
$(0.6, 0.3)$	15	Acc	52.83 ±1.04	55.17 ±0.62	57.92 ±1.45	71.71 ±0.50	81.67 ±0.75	95.83 ±0.31	98.79 ±0.47	99.88 ±0.00
		AUC	53.90 ±0.16	56.23 ±1.06	60.97 ±0.60	79.46 ±0.25	89.15 ±0.33	99.35 ±0.13	99.96 ±0.02	100.00 ±0.00
	25	Acc	50.63 ±1.39	52.63 ±2.13	53.66 ±2.15	62.71 ±1.96	69.08 ±1.56	82.04 ±1.41	92.29 ±0.29	99.04 ±0.19
		AUC	50.79 ±1.58	54.50 ±2.57	54.22 ±0.10	68.33 ±2.26	77.00 ±0.91	89.97 ±0.93	97.87 ±0.11	99.96 ±0.02
	40	Acc	50.44 ±0.92	52.55 ±0.63	50.33 ±2.78	53.44 ±1.27	60.95 ±1.55	74.00 ±0.72	85.78 ±1.57	97.67 ±0.17
		AUC	52.63 ±1.35	54.06 ±0.90	51.75 ±3.28	55.05 ±0.49	66.85 ±0.89	81.66 ±1.15	93.84 ±0.56	99.79 ±0.05
	60	Acc	49.67 ±0.76	49.08 ±2.63	49.58 ±1.04	51.92 ±0.88	50.83 ±0.38	54.75 ±1.75	68.92 ±2.04	83.58 ±1.66
		AUC	49.63 ±1.07	50.46 ±1.61	51.19 ±2.06	53.46 ±0.89	50.39 ±1.52	58.05 ±1.45	73.98 ±2.08	91.26 ±0.80
$(0.8, 0.3)$	15	Acc	51.92 ±0.41	53.12 ±1.84	49.88 ±0.63	55.42 ±0.72	56.21 ±0.26	68.17 ±0.76	84.29 ±0.31	98.37 ±0.21
		AUC	51.80 ±1.20	53.76 ±1.78	50.76 ±1.07	57.72 ±0.91	58.51 ±0.68	75.85 ±0.48	91.70 ±0.23	99.88 ±0.02
	25	Acc	50.04 ±1.00	51.17 ±1.98	48.67 ±3.26	51.42 ±0.64	52.50 ±1.62	56.12 ±2.17	65.91 ±0.32	85.75 ±0.55
		AUC	48.55 ±0.49	52.41 ±1.46	48.30 ±2.61	52.63 ±0.51	53.89 ±1.14	58.44 ±3.40	72.87 ±0.42	93.63 ±0.51
	40	Acc	49.89 ±0.67	49.17 ±0.34	51.17 ±1.00	53.39 ±0.26	50.72 ±2.43	50.67 ±1.00	51.78 ±0.69	53.28 ±0.68
		AUC	50.15 ±0.95	48.90 ±1.40	51.58 ±1.12	55.21 ±0.28	50.50 ±1.39	52.14 ±1.65	51.70 ±0.44	54.80 ±0.55
	60	Acc	48.58 ±0.95	47.67 ±2.96	51.83 ±2.27	47.67 ±2.13	51.75 ±1.56	49.00 ±2.88	47.75 ±2.88	52.92 ±1.18
		AUC	47.61 ±1.22	48.58 ±1.10	51.01 ±0.73	47.33 ±1.58	51.78 ±1.49	48.76 ±3.90	45.12 ±2.71	54.23 ±1.31
$(0.6, 0.5)$	15	Acc	53.13 ±2.14	55.75 ±0.66	59.79 ±1.20	66.50 ±0.13	77.08 ±0.50	89.29 ±0.69	95.38 ±0.76	99.50 ±0.21
		AUC	53.95 ±1.08	58.77 ±0.70	63.55 ±0.89	73.34 ±0.47	85.80 ±0.48	95.85 ±0.19	99.29 ±0.26	99.99 ±0.01
	25	Acc	47.87 ±0.65	52.83 ±0.56	51.12 ±0.38	57.42 ±1.06	60.92 ±0.44	74.92 ±1.70	84.42 ±1.38	94.71 ±0.89
		AUC	47.92 ±2.81	53.88 ±1.42	51.40 ±1.03	60.68 ±0.56	65.71 ±0.99	82.38 ±0.76	92.62 ±1.17	99.16 ±0.19
	40	Acc	55.17 ±0.73	50.39 ±0.85	51.22 ±2.62	51.39 ±1.40	56.33 ±2.59	63.89 ±2.20	74.72 ±0.95	91.61 ±0.51
		AUC	56.78 ±2.35	50.55 ±1.77	51.10 ±1.81	52.12 ±1.46	59.29 ±2.23	68.47 ±1.97	84.51 ±0.55	97.25 ±0.15
	60	Acc	46.25 ±1.50	50.58 ±2.24	49.50 ±3.31	50.00 ±2.14	50.50 ±2.41	55.92 ±1.04	61.08 ±3.71	68.00 ±1.32
		AUC	43.89 ±0.88	49.92 ±2.93	48.99 ±1.16	50.90 ±4.31	51.61 ±2.85	57.36 ±1.23	66.24 ±4.52	74.90 ±2.35
$(0.8, 0.5)$	15	Acc	50.37 ±1.75	51.37 ±0.55	51.42 ±0.72	51.79 ±1.23	52.96 ±0.48	59.92 ±3.50	70.63 ±1.09	89.88 ±0.45
		AUC	49.43 ±1.21	50.59 ±0.62	52.65 ±0.74	53.12 ±0.90	53.95 ±0.93	63.85 ±2.87	78.08 ±0.30	96.46 ±0.22
	25	Acc	48.54 ±1.37	51.33 ±1.69	49.62 ±0.87	50.92 ±1.66	51.45 ±0.76	52.21 ±1.04	55.63 ±0.33	70.29 ±1.04
		AUC	47.85 ±1.23	52.71 ±1.49	49.09 ±0.39	51.16 ±1.54	53.16 ±1.38	52.57 ±2.52	56.96 ±1.05	77.68 ±1.30
	40	Acc	50.11 ±1.00	50.17 ±1.34	49.94 ±0.84	49.22 ±0.96	50.11 ±3.26	49.56 ±2.55	51.67 ±1.04	52.28 ±1.67
		AUC	50.91 ±2.93	49.30 ±1.56	49.48 ±0.43	48.29 ±0.32	50.73 ±3.77	48.59 ±2.35	53.10 ±1.61	54.05 ±2.18
	60	Acc	50.75 ±3.38	52.33 ±1.38	48.58 ±2.31	49.25 ±0.66	49.58 ±0.52	51.83 ±1.76	48.58 ±1.04	47.42 ±0.38
		AUC	50.98 ±3.60	51.68 ±1.19	48.02 ±3.34	49.34 ±0.83	49.76 ±1.71	51.64 ±2.05	47.73 ±0.30	49.07 ±0.95

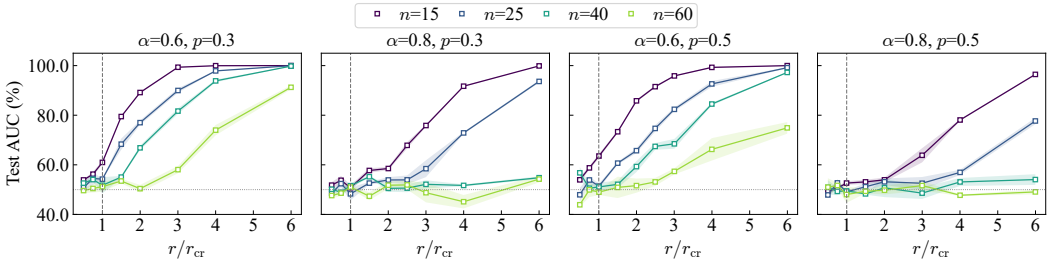


Figure 3: Test AUC of GIN against density normalized by the satisfiability threshold. Curves show means over three seeds, with one standard deviation.

Each class contains 1000–2000 fresh instances, partitioned into disjoint 60/20/20 splits. Models are trained for 200 epochs with Adam, selected by validation accuracy, and evaluated over three seeds. Controls vary the backbone, communication, representation order, and input interface on matched instances and splits. We report accuracy and AUC, using AUC across methods because the pair score has no intrinsic threshold (Hanley & McNeil, 1982; Fawcett, 2006). RQ3 defines *hard* cells by a reference anonymous-graph GIN below 60 AUC. Further details are in Appendix D.

4.1 Size-dependent graph-learning crossover (RQ1)

To examine how the density required for graph classification changes with instance size, we plot the test AUC of a representative message-passing backbone across the normalized density in Figure 3.

We observe a size-dependent crossover as constraints accumulate in Table 1. At $(\alpha, p) = (0.6, 0.5)$, the density reaching 0.70 AUC moves from 1.15 to 1.89, 2.68, and 4.21 times r_{cr} for $n = 15, 25, 40, 60$, and at $(0.8, 0.5)$ it exceeds the sweep for $n \geq 40$. Normalization by r_{cr} does not collapse the curves, separating the satisfiability and empirical learning boundaries. The complete

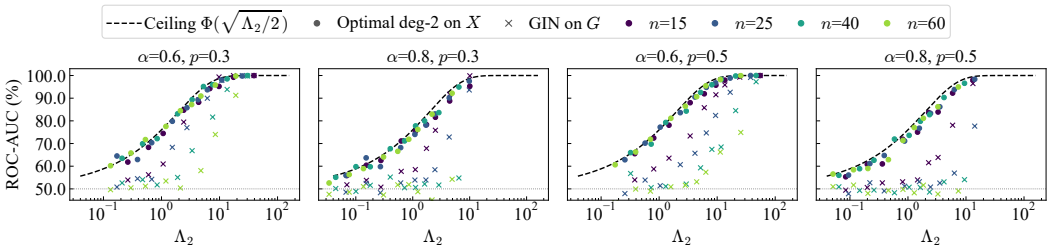


Figure 4: AUC against the second-order signal Λ_2 on a logarithmic scale.

grid in Table 1 shows the same rightward movement across parameter settings. We further vary the crossing definition and message-passing backbone in Appendix E.1 and E.7, where the trend persists for alternative performance levels and for GCN, GraphSAGE, and GAT. These observations are consistent with the theoretical picture in which larger instances require denser overlap before local evidence reveals the global distinction.

4.2 Pairwise visibility follows Λ_2 (RQ2)

We test whether Λ_2 organizes finite-sample pairwise evidence by plotting the shared-variable score and GIN on the same theoretical axis in Figure 4. The shared-variable scores collapse across all four grids, whereas GIN retains a strong size dependence. Under a Gaussian approximation, the exact quadratic SNR predicts $\mathrm{AUC}_{\mathrm{Gauss}}(\Lambda_2) = \Phi(\sqrt{\Lambda_2/2})$. Measured pair-score AUC differs from this reference by 1.35–2.70 points on average, while independent local planting and label shuffling remain near chance. In the descriptive bin $\Lambda_2 \geq 3$, the pair score reaches 91.86 AUC while anonymous-graph GIN remains at 55.08, exposing available but unextracted signal. Appendix E.2 and E.3 report calibration, stratification, and small-n artifact controls.

4.3 Recovering unextracted pairwise evidence (RQ3)

We track the *empirical extraction gap*, defined as pair-score AUC minus anonymous-graph GIN AUC, across density. Figure 5 shows that the gap is generally hump-shaped. For $(\alpha, p) = (0.6, 0.5)$, its maximum grows from 15.86 points at $n = 15$ to 38.81 at $n = 60$ and shifts from $1.5r_{\mathrm{cr}}$ to $2.5r_{\mathrm{cr}}$. We first test whether the gap merely reflects a failure to fit the training data.

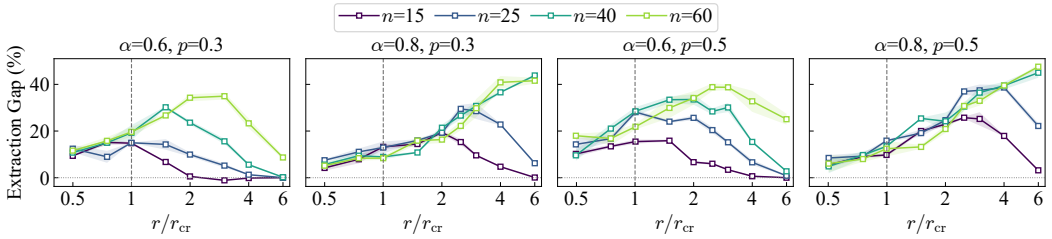


Figure 5: AUC difference between the deduplicated relation-level pair score and GIN along each density sweep.

Figure 6 shows this contrast for $p = 0.3$, while Appendix Figure E.5 reports the complementary $p = 0.5$ grids. Across the two figures, training and test accuracy rise together in easier small-instance cells. In harder large-instance cells, however, GIN often fits substantially more training labels while its held-out accuracy remains near chance. The separation is especially persistent for $\alpha = 0.8$ and for the larger n values, where increasing density can retain high training accuracy without producing a corresponding test improvement. Thus, an inability to reduce training error is not the sole explanation for the extraction gap. The data-scale control in Appendix E.4 supports the same interpretation: an eightfold increase in training data gains 26.33 AUC points in a learnable positive-control cell, but only 6.19 and 12.17 points in two hard cells. The model responds to additional data once the discriminative relation is within reach, whereas more samples alone do not close the hard-cell gap.

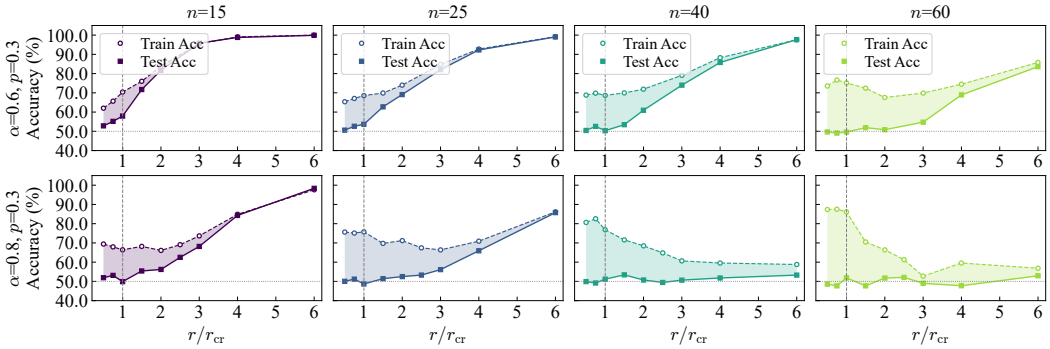


Figure 6: GIN training accuracy (dashed, open markers) and test accuracy (solid, filled markers) across density for the two $p = 0.3$ grids, with shaded regions showing the train–test gap.

To separate the information exposed by the representation from the expressive power of the learner, we vary these two axes independently on the same hard cells. The encoding ladder fixes GIN and compares the anonymous graph, generic structural features, a factor graph, and relation-aligned features. The architecture ladder instead fixes the anonymous encoding and compares a statistics MLP, GIN, GIN with a virtual node, and a Subgraph GNN.

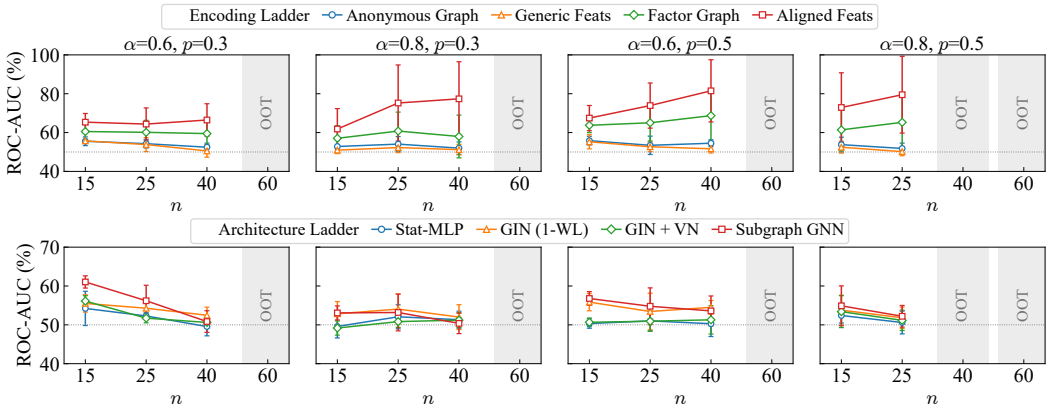


Figure 7: Encoding and architecture ladders on hard cells. The top row fixes GIN while changing the representation, and the bottom row fixes the anonymous encoding while changing the learner. OOT denotes sweeps that exceeded the compute budget.

The upper row of Figure 7 shows a clear alignment pattern. Generic structural features provide little improvement, the factor graph is intermediate, and relation-aligned features yield the strongest recovery. The lower row shows that changing the learner at a fixed anonymous encoding gives smaller and less consistent gains that diminish with size. Within the evaluated interventions, preserving shared-variable relations is therefore more effective than adding generic structural features or model capacity. Matched-order detectors in Appendix E.6 support the same conclusion, while Appendix E.4 reports the data-scale and optimization controls.

5 Conclusion

In this paper, we study when a global distinction becomes learnable from individually uninformative local observations using planted Model RB. Our analysis identifies Λ_2 as the optimal quadratic SNR and proves a local-global separation that widens with problem size. The experiments reveal a steeper empirical learning crossover: shared-variable evidence becomes visible before the evaluated GNNs extract it, while representations preserving these relations recover more signal than the tested generic structural and architectural changes. Learning global distinctions from local observations therefore requires sufficient overlap to make the evidence visible and a representation that preserves the interactions carrying it. We hope this local-global perspective informs the design of future graph learners that incorporate global structural summaries before message passing or jointly model local observations and the global relations organizing them.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and textual revision. The authors reviewed the resulting text and take responsibility for the technical content, experiments, citations, and conclusions of this work.

ETHICS STATEMENT

This work studies theoretically defined random constraint-satisfaction models and uses only synthetically generated instances. It does not involve human participants, personal data, or sensitive real-world datasets. We are not aware of specific ethical risks arising from the experiments or their intended use.

REPRODUCIBILITY STATEMENT

Section 2 specifies the finite Model RB construction and observation model, while Section 3 states the theoretical results. Appendices A–C provide complete notation, background, and proofs. Appendix D documents instance generation, graph encodings, baselines, training, and evaluation; Appendix E reports the complete numerical results and controls. Appendix F gives the reproduction details for the introductory case study.

REFERENCES

Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 793–802, 2008. doi: 10.1109/FOCS.2008.11.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.05205.

Waïss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lxHgXYN4bwl.

Jinheon Baek, Minki Kang, and Sung Ju Hwang. Accurate learning of graph representations with graph multiset pooling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=JHcqXGaqiGn.

Pablo Barceló, Floris Geerts, Juan Reutter, and Maksimilian Ryschkov. Graph neural networks with local graph parameters. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d4d8d1ac7e00e9105775a6b660dd3cbb-Abstract.html.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks, 2018. URL https://arxiv.org/abs/1806.01261.

Beatrice Bevilacqua, Fabrizio Frasca, Derek Lim, Balasubramaniam Srinivasan, Chen Cai, Gopinath Balamurugan, Michael M. Bronstein, and Haggai Maron. Equivariant subgraph aggregation networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dFbKQaRk15w.

Cristian Bodnar, Fabrizio Frasca, Nina Otter, Yu Guang Wang, Pietro Liò, Guido F. Montúfar, and Michael M. Bronstein. Weisfeiler and lehman go cellular: CW networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2625–2640, 2021.

Jan Böker, Ron Levie, Ningyuan Huang, Soledad Villar, and Christopher Morris. Fine-grained expressivity of graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21527–21543, 2023.

Giorgos Bouritsas, Fabrizio Frasca, Stefanos Zafeiriou, and Michael M. Bronstein. Improving graph neural network expressivity via subgraph isomorphism counting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):657–668, 2023. doi: 10.1109/TPAMI.2022.3154319.

Matthew S. Brennan, Guy Bresler, Samuel B. Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 774–774, 2021. URL https://proceedings.mlr.press/v134/brennan21a.html.

Jin-Yi Cai, Martin Fürer, and Neil Immerman. An optimal lower bound on the number of variables for graph identification. *Combinatorica*, 12(4):389–410, 1992. doi: 10.1007/BF01305232.

Peter Cheeseman, Bob Kanefsky, and William M. Taylor. Where the really hard problems are. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, pp. 331–337, 1991.

Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? In *Advances in Neural Information Processing Systems*, volume 33, pp. 10383–10395, 2020.

Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Veličković. Principal neighbourhood aggregation for graph nets. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13260–13271, 2020.

Zdeněk Dvořák. On recognizing graphs by numbers of homomorphisms. *Journal of Graph Theory*, 64(4):330–342, 2010. doi: 10.1002/jgt.20461.

Vijay Prakash Dwivedi, Ladislav Rampášek, Mikhail Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. LRGB: A benchmark for learning on long-range graphs. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010. URL https://doi.org/10.1016/j.patrec.2005.10.010.

Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 77–86, 2015. doi: 10.1145/2746539.2746621.

Fabrizio Frasca, Beatrice Bevilacqua, Michael M. Bronstein, and Haggai Maron. Understanding and extending subgraph GNNs by rethinking their symmetries. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cb2a4cc70db72ea779abd01107782c7b-Abstract-Conference.html.

Ehud Friedgut. Sharp thresholds of graph properties, and the k-SAT problem. *Journal of the American Mathematical Society*, 12(4):1017–1054, 1999. doi: 10.1090/S0894-0347-99-00305-7.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *NeurIPS*, 30, 2017.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747. URL https://doi.org/10.1148/radiology.143.1.7063747.

Ningyuan Huang and Soledad Villar. A short tutorial on the weisfeiler–lehman test and its variants. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8533–8537, 2021. doi: 10.1109/ICASSP39728.2021.9413523.

Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21618–21629, 2021.

Florent Krzakala and Lenka Zdeborová. Hiding quiet solutions in random constraint satisfaction problems. *Physical Review Letters*, 102(23):238701, 2009. doi: 10.1103/PhysRevLett.102.238701.

Dmitriy Kunisky, Alexander S. Wein, and Afonso S. Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio, 2019. URL https://arxiv.org/abs/1907.11636.

Angsheng Li and Yicheng Pan. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6):3290–3339, 2016.

Angsheng Li, Xianchen Yin, Bingxiang Xu, Danyang Wang, Jimin Han, Yi Wei, Yun Deng, Ying Xiong, and Zhihua Zhang. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Communications*, 9(1):3265, 2018.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05493.

Andreas Loukas. What graph neural networks cannot learn: Depth vs width. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=B1l2bp4YwS.

Andreas Loukas. How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3465–3476, 2020b.

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2153–2164, 2019a. URL https://proceedings.neurips.cc/paper/2019/hash/bb04af0f7ecaee4aae62035497da1387-Abstract.html.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=Syx72jC9tm.

Sohir Maskey, Raffaele Paolino, Fabian Jogl, Gitta Kutyniok, and Johannes F. Lutzeyer. Graph representational learning: When does more expressivity hurt generalization? In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C6vpifaZvU.

David Mitchell, Bart Selman, and Hector Levesque. Hard and easy distributions of SAT problems. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pp. 459–465, 1992.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4602–4609, 2019. doi: 10.1609/aaai.v33i01.33014602. URL https://doi.org/10.1609/aaai.v33i01.33014602.

Christopher Morris, Gaurav Rattan, and Petra Mutzel. Weisfeiler and leman go sparse: Towards scalable higher-order graph embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21824–21840, 2020.

Ryan L. Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4663–4673, 2019.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14501–14515, 2022.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkx1qkrKPr.

Ryoma Sato. A survey on the expressive power of graph neural networks, 2020. URL https://arxiv.org/abs/2003.04078.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler–lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7UmjRGzp-A.

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

Clément Vignac, Andreas Loukas, and Pascal Frossard. Building powerful and equivariant graph neural networks with structural message-passing. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14143–14155, 2020.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871, 2019.

Junran Wu, Xueyuan Chen, Ke Xu, and Shangzhe Li. Structural entropy guided graph hierarchical pooling. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 24017–24030, 2022.

Ke Xu and Wei Li. Exact phase transitions in random constraint satisfaction problems. *Journal of Artificial Intelligence Research*, 12:93–103, 2000. doi: 10.1613/JAIR.696. URL https://doi.org/10.1613/jair.696.

Ke Xu, Frédéric Boussemart, Fred Hemery, and Christophe Lecoutre. Random constraint satisfaction: Easy generation of hard (satisfiable) instances. *Artificial Intelligence*, 171(8–9):514–534, 2007. doi: 10.1016/j.artint.2007.04.001. URL https://doi.org/10.1016/j.artint.2007.04.001.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5453–5462, 2018.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bohang Zhang, Jingchu Gai, Yiheng Du, Qiwei Ye, Di He, and Liwei Wang. Beyond Weisfeiler–Lehman: A quantitative framework for GNN expressiveness. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ec702dd6e83b2113a43614685a7e2ac6-Abstract-Conference.html.

Muhan Zhang and Pan Li. Nested graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15734–15747, 2021.

Lingxiao Zhao and Leman Akoglu. PairNorm: Tackling oversmoothing in GNNs. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkecl1rtwB.

Appendix

Table of Contents

A Notation **16**

B Model RB Background and Finite Conventions **16**
B.1 Original and realized parameters 16
B.2 Classical satisfiability threshold 17
B.3 Forced satisfiability as size biasing 17
B.4 An exact negative control 18

C Technical Proofs and Derivations **18**
C.1 Likelihood ratio under deterministic encoding 18
C.2 Likelihood-ratio identities 19
C.3 The centered relation kernel 19
C.4 Orthogonal block expansion and exact low-degree norm 20
C.5 Exact second-order norm and variational identity 21
C.6 Finite arity-two formula 22
C.7 Asymptotic low-degree density 22
C.8 Logarithmic moments of the overlap kernel 23
C.9 Uniform logarithmic-degree bound 24
C.10 Local–global separation 25

D Experimental Details **25**
D.1 Instance generation 25
D.2 Microstructure graph and observation interfaces 25
D.3 Baseline details 26
D.4 Pair score and relation-information ladder 26
D.5 Node features for the encoding ladder 27
D.6 Negative and shallow controls 28
D.7 Training and evaluation protocol 29
D.8 Experimental setting 29

E Extended Experimental Results **29**
E.1 Sensitivity of the empirical crossing 29
E.2 Calibration residual against the Gaussian reference 31
E.3 Shallow graph summaries 32
E.4 Controls on the extraction gap 32
E.5 Communication range and representation order 36
E.6 Disentangling representation and architecture 36
E.7 Backbone sweep, resolved by size and density 38
E.8 Crossings beyond the swept density range 39

F Reproducibility Details for the Introductory Case Study **40**
F.1 Configuration and sampling 40
F.2 Label verification and reported statistics 41

G Extended Related Work **41**

H Implications for Local-to-Global Graph Learning **43**

A NOTATION

Table A.1 collects the symbols used throughout the paper. The parameters n and m count CSP variables and constraints, respectively. By contrast, the encoded graph $G = (\mathcal{V}, \mathcal{E}) = \phi(X)$ contains $|\mathcal{V}| = nq$ value nodes and an instance-dependent number of edges $|\mathcal{E}|$. Thus $n \neq |\mathcal{V}|$ and $m \neq |\mathcal{E}|$. For example, when $\alpha = 0.6$ and $n = 60$, the CSP has 60 variables and the graph has 720 nodes. At density $r = 2$, the CSP has approximately 492 constraints, each contributing multiple graph edges. We use n, m for CSP size and $|\mathcal{V}|, |\mathcal{E}|$ for graph size.

Table A.1: Summary of notation and descriptions. The upper block is the constraint-satisfaction instance, the middle block is the detection problem and its low-degree analysis, and the lower block is the graph encoding and the learned models.

Notation	Description
n	Number of CSP **variables** (not the number of graph nodes)
k	Constraint arity, with $k = 2$ throughout the experiments
q	Realized integer domain size per variable, $q = \mathrm{round}(n^\alpha)$
α	Domain exponent, $q = n^\alpha$, with $\alpha > 1/k$ required by the classical threshold
m	Number of **constraints** (not the number of graph edges), $m = \mathrm{round}(rn \log n)$
r	Constraint density and the primary control parameter of every sweep
p	Nominal constraint tightness (fraction of forbidden tuples)
s, ρ, p_{eff}	Allowed tuples per table, $\rho = s/N$, and realized tightness $p_{\text{eff}} = 1 - \rho$
ω	Tightness odds, $\omega = p_{\text{eff}}/(1 - p_{\text{eff}})$
N	Number of local tuples per constraint, $N = q^k$
S_a, R_a	Scope and allowed relation table of constraint a
$X = (S, R)$	The complete observation: all scopes together with all relation tables
σ^*	Hidden planted assignment, $\sigma^* \in [q]^n$
$Z(X)$	Number of satisfying assignments of the instance X
r_{cr}	Classical satisfiability threshold, $r_{\text{cr}} = \alpha/[-\log(1-p)]$
$\mathbb{P}_0$, $\mathbb{P}_1$	Null and mixed planted distributions over complete observations
$\Pr(A)$	Probability of an event A under the random law specified in context
L_X	Likelihood ratio $\mathrm{d}\mathbb{P}_1/\mathrm{d}\mathbb{P}_0$ on the complete observation
$Y_{a,z}$	Indicator that tuple z is allowed by constraint a
D	Truncation degree of the low-degree projection
$\mathcal{H}_{\leq D}$	Fiberwise space of relation-table polynomials of degree at most D
$L_{\leq D}$	Degree-D orthogonal projection of L_X, taken fiberwise over scopes
J	Overlap size of two independent random scopes, $J = \lvert S_a \cap S_b \rvert$
Λ_2	Exact second-order signal: excess degree-two norm and optimal squared SNR
$r_{2,*} = r_{2,1}$	Density at which $\Lambda_2 = 1$, the theoretical unit-SNR reference
$r_{\text{LD}}(n)$	Asymptotic low-degree visibility scale
$\Phi(\cdot)$	Standard normal distribution function
AUC	Area under the ROC curve, reported as a percentage in tables
ϕ, $G = (\mathcal{V}, \mathcal{E}) = \phi(X)$	Deterministic graph encoding and the resulting graph
$\lvert\mathcal{V}\rvert$, $\lvert\mathcal{E}\rvert$	Number of **graph nodes** ($\lvert\mathcal{V}\rvert = nq$) and **graph edges**
L_G	Push-forward likelihood ratio on the graph, $L_G = \mathbb{E}_0[L_X \mid G]$
L_0, L_1, L_2	Relation-information levels: raw, scope-grouped, and tuple-deduplicated
r_G	AUC crossing of the deduplicated relation-level pair score (legacy plot notation)
r_{GIN}	Empirical crossing of the trained model at the same level
h_{GNN}	Number of message-passing layers (network depth)

B MODEL RB BACKGROUND AND FINITE CONVENTIONS

This appendix summarizes the original Model RB and forced-satisfiable constructions, separating classical properties of the ensemble from the low-degree results developed in this paper.

B.1 ORIGINAL AND REALIZED PARAMETERS

Lemma B.1 (Rounding preserves the asymptotic parameters)**.** Fix $\alpha > 0$, $p \in (0, 1)$, $r > 0$, and fixed arity k. Let

$$q_n = \mathrm{round}(n^\alpha), \qquad t_n = \mathrm{round}(pq_n^k), \qquad m_n = \mathrm{round}(rn \log n), \tag{B.1}$$

and define $p_{\text{eff},n} = t_n/q_n^k$. Then

$$\log q_n = \alpha \log n + o(1), \qquad p_{\text{eff},n} = p + O(q_n^{-k}), \qquad m_n = rn \log n + O(1). \tag{B.2}$$

In particular, every exact finite formula in the paper may be evaluated at $(q_n, m_n, p_{\mathrm{eff},n})$, while its stated Model RB asymptotics retain the nominal parameters (α, r, p).

Proof. The rounding error of each displayed integer is at most $1/2$. Since $q_n = n^\alpha(1 + O(n^{-\alpha}))$, taking logarithms gives the first claim. Dividing $|t_n - pq_n^k| \le 1/2$ by q_n^k gives the second, and the third is immediate. □

The original formulation draws $m = rn \log n$ scopes independently, with distinct variables inside a scope and repetitions permitted across scopes, and selects exactly pq^k incompatible tuples without replacement in each relation (Xu & Li, 2000). Thus our use of a fixed-cardinality relation and repeated scopes follows Model RB rather than an independent-Bernoulli approximation. Rounding finite quantities is also the convention used in the experimental Model RB literature (Xu et al., 2007).

B.2 Classical satisfiability threshold

Theorem B.2 (Exact Model RB threshold, Xu & Li, 2000)**.** Suppose $q = n^\alpha$, $m = rn \log n$, $\alpha > 1/k$, and $k \ge 1/(1-p)$, with fixed k, α, and $p \in (0,1)$. Let

$$r_{\mathrm{cr}} = \frac{\alpha}{-\log(1-p)}. \tag{B.3}$$

For fixed $r < r_{\mathrm{cr}}$, a Model RB instance is satisfiable with probability tending to one. For fixed $r > r_{\mathrm{cr}}$, it is unsatisfiable with probability tending to one.

Proof. Write Z for the number of satisfying assignments. Under the null model,

$$\mathbb{E}_0 Z = q^n (1-p)^m = \exp\{n \log n \, [\alpha + r \log(1-p)]\}. \tag{B.4}$$

For $r > r_{\mathrm{cr}}$ this expectation tends to zero, and Markov's inequality proves unsatisfiability with high probability. For $r < r_{\mathrm{cr}}$, Xu & Li (2000) evaluate the assignment-overlap sum in the second moment and prove

$$\frac{\mathbb{E}_0 Z^2}{(\mathbb{E}_0 Z)^2} \longrightarrow 1 \tag{B.5}$$

under the stated conditions. The Paley–Zygmund inequality then gives $\mathbb{P}_0(Z > 0) \to 1$. □

Remark B.1 (Distributional and instance-level statements)**.** Theorem B.2 describes the probability of satisfiability for a random instance sequence. Determining whether a particular realized null instance is satisfiable remains an instance-level computation. In our testing problem, labels record the known generating distribution: $\mathbb{P}_0$ for a null draw and $\mathbb{P}_1$ for a planted draw. Above r_{cr}, these generative labels agree with unsatisfiable versus satisfiable with probability tending to one because a planted draw always contains its witness and a null draw is asymptotically unsatisfiable. Below r_{cr}, both classes can be satisfiable, so detection, search, and satisfiability remain distinct tasks.

B.3 Forced satisfiability as size biasing

The forced-satisfiable Model RB generator of Xu et al. (2007) first draws an assignment and then chooses every relation uniformly subject to containing the induced local tuple. This is exactly the mixed planted distribution used in this paper.

Proposition B.3 (Size-bias identity for forced Model RB)**.** For every integrable statistic $h(X)$,

$$\mathbb{E}_1 h(X) = \frac{\mathbb{E}_0[Z(X)h(X)]}{\mathbb{E}_0 Z(X)}. \tag{B.6}$$

Consequently,

$$\mathbb{E}_1 Z = \frac{\mathbb{E}_0 Z^2}{\mathbb{E}_0 Z}. \tag{B.7}$$

Under the conditions of Theorem B.2 with $r < r_{\mathrm{cr}}$,

$$\frac{\mathbb{E}_1 Z}{\mathbb{E}_0 Z} \longrightarrow 1. \tag{B.8}$$

Proof. Theorem 3.1 gives $\mathrm{d}\mathbb{P}_1/\,\mathrm{d}\mathbb{P}_0 = Z/\mathbb{E}_0 Z$. Multiplying by h and integrating proves the first identity. Setting $h = Z$ proves Eq. (B.7). The final claim follows from the second-moment limit used in Theorem B.2. This recovers the forced-solution identity highlighted by Xu et al. (2007). □

Remark B.2 (Connection to the detection problem)**.** The threshold theorem concerns global satisfiability, whereas Proposition B.3 explains the distributional meaning of planting. Neither result determines the low-degree detection scale. Our contribution begins by projecting this size-biased likelihood ratio and asking which interaction orders can see it.

B.4 An exact negative control

Lemma B.4 (Independent local planting equals the null law)**.** For each constraint independently, draw $U_a \sim \mathrm{Unif}([q]^k)$ and then draw R_a uniformly among size-s relations containing U_a. After marginalizing $(U_a)_{a=1}^m$, the joint law of the relation tables, conditional on the scopes, is exactly the null law.

Proof. For any fixed size-s relation R,

$$\Pr(R_a = R) = \sum_{u \in R} \frac{1}{N} \binom{N-1}{s-1}^{-1} = \binom{N}{s}^{-1}. \tag{B.9}$$

The independently sampled local tuples also preserve independence across constraints. Hence the full conditional distribution is $\mathbb{P}_0$. □

This control matches the one-constraint planting mechanism but removes the single global assignment that correlates different constraints. Its population AUC is therefore exactly $1/2$ for every statistic. Finite-sample deviations from $1/2$ quantify only estimation variability and model selection.

C Technical Proofs and Derivations

Each result is restated before its proof with an appendix-local number and a reference to the corresponding result in the main text.

C.1 Likelihood ratio under deterministic encoding

Proposition C.1 (Likelihood ratio after deterministic encoding (restatement of Proposition 2.5))**.** Assume $\mathbb{P}_1 \ll \mathbb{P}_0$ and let $G = \phi(X)$ be a deterministic encoding. Then

$$L_G(G) = \frac{\mathrm{d}\mathbb{P}_1^G}{\mathrm{d}\mathbb{P}_0^G}(G) = \mathbb{E}_0[L_X(X) \mid G], \tag{C.1}$$

and

$$\chi^2(\mathbb{P}_1^G \| \mathbb{P}_0^G) \le \chi^2(\mathbb{P}_1 \| \mathbb{P}_0). \tag{C.2}$$

Proof. For every bounded measurable h,

$$\mathbb{E}_{\mathbb{P}_1^G}[h(G)] = \mathbb{E}_{\mathbb{P}_1}[h(\phi(X))] = \mathbb{E}_{\mathbb{P}_0}[L_X(X) h(\phi(X))] \tag{C.3}$$

$$= \mathbb{E}_{\mathbb{P}_0}[\mathbb{E}_0[L_X(X) \mid G] h(G)]\,. \tag{C.4}$$

This proves the likelihood-ratio identity. Since conditional expectation is an $L^2(\mathbb{P}_0)$ contraction,

$$\mathbb{E}_0[(L_G - 1)^2] \le \mathbb{E}_0[(L_X - 1)^2], \tag{C.5}$$

which is the stated chi-square contraction. □

C.2 LIKELIHOOD-RATIO IDENTITIES

Lemma C.2 (One-constraint likelihood ratio)**.** Let $\mathbb{P}_{1,x}$ be the uniform distribution over size-s relations containing $x \in [q]^k$, and let

$$\ell_x(R) := \frac{\mathbf{1}\{x \in R\}}{\rho}. \tag{C.6}$$

Then

$$\frac{\mathbb{P}_{1,x}(R)}{\mathbb{P}_0(R)} = \ell_x(R). \tag{C.7}$$

Proof. Under $\mathbb{P}_0$, every size-s relation has mass $\binom{N}{s}^{-1}$. Under $\mathbb{P}_{1,x}$, a relation has mass $\binom{N-1}{s-1}^{-1}$ if it contains x and zero otherwise. Therefore

$$\frac{\mathbb{P}_{1,x}(R)}{\mathbb{P}_0(R)} = \mathbf{1}\{x \in R\}\frac{\binom{N}{s}}{\binom{N-1}{s-1}} = \frac{\mathbf{1}\{x \in R\}}{s/N}. \tag{C.8}$$

□

Theorem C.3 (Likelihood ratio equals normalized solution count (restatement of Theorem 3.1))**.** For the mixed planted and null distributions of Section 2, the likelihood ratio of the complete observation satisfies

$$L_X(X) = \frac{Z(X)}{q^n\rho^m} = \frac{Z(X)}{\mathbb{E}_0 Z(X)}. \tag{C.9}$$

Proof. Fix $\sigma \in [q]^n$. Conditional on S, independence of the relation tables gives

$$L_\sigma(X) = \prod_{a=1}^{m} \frac{\mathbf{1}\{\sigma_{S_a} \in R_a\}}{\rho}. \tag{C.10}$$

Marginalizing the uniform hidden assignment yields

$$L_X(X) = \frac{1}{q^n}\sum_{\sigma\in[q]^n} L_\sigma(X) \tag{C.11}$$

$$= \frac{1}{q^n\rho^m}\sum_{\sigma\in[q]^n}\prod_{a=1}^{m}\mathbf{1}\{\sigma_{S_a} \in R_a\} \tag{C.12}$$

$$= \frac{Z(X)}{q^n\rho^m}. \tag{C.13}$$

Under $\mathbb{P}_0$, a fixed assignment satisfies each independent relation with probability ρ, so $\mathbb{E}_0 Z = q^n\rho^m$. □

C.3 THE CENTERED RELATION KERNEL

Lemma C.4 (Centered one-constraint kernel)**.** Let $g_x(R) := \ell_x(R) - 1$ and $N = q^k$. For every $x, y \in [q]^k$,

$$c(x, y) := \mathbb{E}_0[g_x(R)g_y(R)] = \omega\frac{N\mathbf{1}\{x = y\} - 1}{N - 1}. \tag{C.14}$$

Proof. If $x = y$, then $\mathbb{P}_0(x \in R) = \rho$, and hence

$$c(x, x) = \mathbb{E}_0\ell_x^2 - 1 = \rho^{-1} - 1 = \omega. \tag{C.15}$$

If $x \neq y$, uniform sampling without replacement gives

$$\mathbb{P}_0(x, y \in R) = \frac{s(s-1)}{N(N-1)}. \tag{C.16}$$

Consequently,

$$c(x, y) = \frac{s(s-1)}{N(N-1)\rho^2} - 1 = -\frac{1-\rho}{\rho(N-1)} = -\frac{\omega}{N-1}, \tag{C.17}$$

which proves Eq. (C.14). □

C.4 Orthogonal block expansion and exact low-degree norm

Lemma C.5 (Block expansion and ordinary polynomial degree)**.** For $A \subseteq [m]$, define

$$L_A(S,R) := \mathbb{E}_\sigma \prod_{a\in A} g_{\sigma_{S_a}}(R_a), \qquad L_\varnothing := 1. \tag{C.18}$$

Then $L_X = \sum_{A\subseteq[m]} L_A$, the family $\{L_A\}$ is pairwise orthogonal in $L^2(\mathbb{P}_0)$, and

$$L_{\le D} = \sum_{|A|\le \min\{D,m\}} L_A. \tag{C.19}$$

The projection is the ordinary coordinate-degree projection defined in Eq. (5).

Proof. For fixed σ, expand

$$\prod_{a=1}^{m} \ell_{\sigma_{S_a}}(R_a) = \prod_{a=1}^{m} \big(1 + g_{\sigma_{S_a}}(R_a)\big) = \sum_{A\subseteq[m]} \prod_{a\in A} g_{\sigma_{S_a}}(R_a). \tag{C.20}$$

Averaging over σ proves $L_X = \sum_A L_A$. If $A \neq B$, choose $a \in A \triangle B$. Conditional on every other relation table and on the scopes, exactly one factor in $L_A L_B$ contains a centered function of R_a. Integrating R_a first gives zero, proving orthogonality.

Each $g_{\sigma_{S_a}}(R_a)$ is linear in the indicators $\{Y_{a,z}\}_{z\in[q]^k}$, so L_A has ordinary polynomial degree at most $|A|$. Conversely, a monomial of degree at most D involves at most D relation-table blocks. If $|A| > D$, at least one centered R_a factor is absent from the monomial, and integration over that table again gives zero. Thus every L_A with $|A| > D$ is orthogonal to the degree-D polynomial space, proving Eq. (C.19). □

Theorem C.6 (Exact low-degree norm (restatement of Theorem 3.2))**.** Let $H \sim \mathrm{Bin}(n, 1/q)$ and

$$\gamma_H = \omega \frac{q^k (H)_k/(n)_k - 1}{q^k - 1}. \tag{C.21}$$

For every integer $D \ge 0$,

$$\|L_{\le D}\|_2^2 = \mathbb{E}_H \left[\sum_{t=0}^{\min\{D,m\}} \binom{m}{t} \gamma_H^t \right]. \tag{C.22}$$

Proof. By Lemma C.5 and orthogonality,

$$\|L_{\le D}\|_2^2 = \sum_{t=0}^{\min\{D,m\}} \sum_{A:|A|=t} \|L_A\|_2^2. \tag{C.23}$$

Fix A with $|A| = t$ and introduce independent assignments σ, τ. Conditional on (σ, τ, S), independence of the relation tables and Lemma C.4 give

$$\|L_A\|_2^2 = \mathbb{E}_{\sigma,\tau,S} \prod_{a\in A} c(\sigma_{S_a}, \tau_{S_a}). \tag{C.24}$$

Conditional on H, a uniform scope lies in the agreement set with probability $(H)_k/(n)_k$. Averaging the centered kernel over one scope gives γ_H. The scopes are independent across a, so the product contributes γ_H^t. Hence $\|L_A\|_2^2 = \mathbb{E}_H \gamma_H^t$. Summing the $\binom{m}{t}$ subsets proves the identity. □

Corollary C.7 (Exact quiet planting (restatement of Corollary 3.3))**.** At every finite n,

$$(L_X)_{=1} = 0, \qquad \|L_{\le 1}\|_2^2 = 1. \tag{C.25}$$

Equivalently, the mixed planted and null distributions have identical one-constraint marginals.

Proof. For $H \sim \mathrm{Bin}(n, 1/q)$,

$$\mathbb{E}(H)_k = (n)_k q^{-k}, \tag{C.26}$$

because $(H)_k$ counts ordered k-tuples of coordinates on which two independent assignments agree. Therefore $\mathbb{E}\gamma_H = 0$. Theorem 3.2 with $D = 1$ gives the norm identity. The degree-one block components are mutually orthogonal, so their total squared norm can vanish only when the degree-one projection is zero. □

C.5 EXACT SECOND-ORDER NORM AND VARIATIONAL IDENTITY

Throughout this subsection, $\Pr(\cdot)$ denotes the probability of an event under the auxiliary random scopes or assignments introduced in the surrounding statement. It is distinct from $\mathbb{P}_0$ and $\mathbb{P}_1$, which denote the two probability measures on complete Model RB observations.

Theorem C.8 (Exact second-order signal (restatement of Theorem 3.4))**.** Let J be the overlap size of two independent uniform scopes and define

$$B_{n,q,k,p_{\text{eff}}} := \frac{\omega^2}{(q^k-1)^2}\sum_{j=1}^{k}\Pr(J=j)(q^j-1), \qquad \Lambda_2 := \binom{m}{2}B_{n,q,k,p_{\text{eff}}}. \tag{C.27}$$

Then

$$\|L_{\le 2}\|_2^2 - 1 = \Lambda_2, \tag{C.28}$$

and

$$\sup_{\substack{f\in\mathcal{H}_{\le 2},\ \mathbb{E}_0 f=0\\ \mathrm{Var}_0(f)>0}} \frac{(\mathbb{E}_1 f - \mathbb{E}_0 f)^2}{\mathrm{Var}_0(f)} = \Lambda_2. \tag{C.29}$$

Consequently, $\sqrt{\Lambda_2}$ is the optimal standardized mean separation among all quadratic detectors on the complete observation.

Proof. Theorem 3.2 and Corollary 3.3 give

$$\|L_{\le 2}\|_2^2 - 1 = \binom{m}{2}\mathbb{E}\gamma_H^2. \tag{C.30}$$

Conditional on $J=j$, let A_S denote the event that two independent assignments σ,τ agree on every coordinate of S, and define A_T analogously. Then

$$\Pr(A_S) = \Pr(A_T) = q^{-k}, \qquad \Pr(A_S\cap A_T \mid J=j) = q^{-(2k-j)}. \tag{C.31}$$

Expanding the two centered kernels gives

$$\mathbb{E}[c(\sigma_S,\tau_S)c(\sigma_T,\tau_T)\mid J=j] \tag{C.32}$$

$$= \frac{\omega^2}{(q^k-1)^2}\mathbb{E}[(q^k\mathbf{1}_{A_S}-1)(q^k\mathbf{1}_{A_T}-1)] = \omega^2\frac{q^j-1}{(q^k-1)^2}. \tag{C.33}$$

The $j=0$ contribution is zero. Averaging over Eq. (9) proves Eq. (12).

For the variational identity, set $S_2^* := (L_X)_{=2}$. For every centered $f\in\mathcal{H}_{\le 2}$,

$$\mathbb{E}_1 f - \mathbb{E}_0 f = \langle L_X - 1, f\rangle_0 = \langle S_2^*, f\rangle_0, \tag{C.34}$$

where the second equality uses orthogonal projection and the vanishing degree-one component. Cauchy–Schwarz yields

$$(\mathbb{E}_1 f - \mathbb{E}_0 f)^2 \le \|S_2^*\|_2^2\,\mathrm{Var}_0(f). \tag{C.35}$$

Equality is attained by every nonzero scalar multiple of S_2^*, and $\|S_2^*\|_2^2 = \Lambda_2$ by Eq. (12). □

Constructive form of the optimal quadratic score. The optimizer used above is not merely existential. Lemma C.5 gives

$$S_2^*(S,R) = \sum_{1\le a<b\le m}\mathbb{E}_\sigma\left[g_{\sigma_{S_a}}(R_a)g_{\sigma_{S_b}}(R_b)\right]. \tag{C.36}$$

If $S_a\cap S_b=\varnothing$, the summand vanishes because the two local tuples are independent and

$$\frac{1}{q^k}\sum_{x\in[q]^k} g_x(R_a) = \frac{|R_a|}{q^k\rho} - 1 = 0. \tag{C.37}$$

Hence only overlapping constraint pairs need to be evaluated. Moreover,

$$\mathbb{E}_0 S_2^* = 0, \qquad \mathrm{Var}_0(S_2^*) = \Lambda_2, \qquad \mathbb{E}_1 S_2^* = \Lambda_2. \tag{C.38}$$

This supplies the explicit oracle pair detector used to connect the theory to experiments. Its null-standardized mean shift is exactly $\sqrt{\Lambda_2}$.

Algorithmic scope. For fixed arity, Eq. (C.36) is a polynomial-time statistic because disjoint scopes contribute zero and each overlapping pair requires only a finite-table contraction. For $k = 2$, the one-variable-overlap term reduces to inner products of q-dimensional centered relation marginals. Grouping those marginals by variable evaluates the sum over all such pairs through the quadratic identity in Eq. (D.2), without materializing a dense state for every pair of graph nodes. Repeated scopes are grouped separately and contribute centered inner products of their full q^2-dimensional tables. The resulting score is an explicit algorithm for testing whether an instance was sampled from the planted or null law. It is not a satisfiability algorithm and does not by itself output a satisfying assignment.

C.6 Finite arity-two formula

Corollary C.9 (Closed form for arity $k = 2$ (restatement of Corollary 3.5))**.** For arity-two constraints and independent scopes sampled with replacement,

$$B_{n,q,2,p_{\text{eff}}} = 2\left(\frac{p_{\text{eff}}}{1-p_{\text{eff}}}\right)^2 \frac{2n+q-3}{n(n-1)(q-1)(q+1)^2}. \tag{C.39}$$

The real-valued constraint count at the unit-SNR contour is

$$m_{2,*} = \frac{1}{2}\left[1+\sqrt{1+4\left(\frac{1-p_{\text{eff}}}{p_{\text{eff}}}\right)^2 \frac{n(n-1)(q-1)(q+1)^2}{2n+q-3}}\right]. \tag{C.40}$$

For an integer experiment, the closest admissible integer contour is obtained by rounding $m_{2,*}$ according to the declared experimental convention.

Proof. For $k = 2$, two independent scopes satisfy

$$\Pr(J=1) = \frac{4(n-2)}{n(n-1)}, \qquad \Pr(J=2) = \frac{2}{n(n-1)}. \tag{C.41}$$

Consequently,

$$\sum_{j=1}^{2} \Pr(J=j)(q^j-1) = \frac{4(n-2)(q-1)+2(q^2-1)}{n(n-1)} \tag{C.42}$$

$$= \frac{2(q-1)(2n+q-3)}{n(n-1)}. \tag{C.43}$$

Substitution into Eq. (10), together with $q^2-1 = (q-1)(q+1)$, yields Eq. (15). The equation $\binom{m}{2}B_{n,q,2,p_{\text{eff}}} = 1$ is quadratic in m. Taking its positive root gives Eq. (C.40). □

C.7 Asymptotic low-degree density

Proposition C.10 (Asymptotic low-degree scale (restatement of Proposition 3.6))**.** Fix $k \geq 2$ and suppose $q = o(n)$ and $p_{\text{eff},n} \to p \in (0,1)$. Whenever $m \to \infty$,

$$\Lambda_2 = \frac{k^2}{2}\left(\frac{p}{1-p}\right)^2 \frac{m^2}{nq^{2k-1}}(1+o(1)). \tag{C.44}$$

The corresponding unit-SNR constraint scale is

$$m_{\text{LD}}(n) \sim \frac{\sqrt{2}(1-p)}{kp}\sqrt{nq^{2k-1}}. \tag{C.45}$$

If $q = n^\alpha$ and $m = rn\log n$, then

$$r_{\text{LD}}(n) \sim \frac{\sqrt{2}(1-p)}{kp}\frac{n^{\alpha(k-1/2)-1/2}}{\log n}. \tag{C.46}$$

Proof. For fixed k and $q = o(n)$, the hypergeometric overlap law gives

$$\Pr(J = 1) = \frac{k^2}{n}(1 + o(1)). \tag{C.47}$$

For each fixed $j \geq 2$, the ratio of the j-overlap contribution in Eq. (10) to the $j = 1$ contribution is

$$\frac{\Pr(J = j)(q^j - 1)}{\Pr(J = 1)(q - 1)} = O\left(\left(\frac{q}{n}\right)^{j-1}\right) = o(1). \tag{C.48}$$

Hence

$$B_{n,q,k,p_{\text{eff}}} = \left(\frac{p}{1-p}\right)^2 \frac{k^2}{nq^{2k-1}}(1 + o(1)). \tag{C.49}$$

Because $m \to \infty$, we have $\binom{m}{2} = m^2(1 + o(1))/2$, which proves Eq. (16). Equating the leading term to one gives Eq. (C.45). Dividing by $n \log n$ and substituting $q = n^\alpha$ gives Eq. (17). □

Corollary C.11 (Arity-two separation from satisfiability (restatement of Corollary 3.7))**.** Suppose $k = 2$, $1/2 < \alpha < 1$, $p \leq 1/2$, and the classical Model RB threshold assumptions hold. Then

$$\frac{r_{\text{LD}}(n)}{r_{\text{cr}}} \sim \frac{(1-p)[-\log(1-p)]}{\sqrt{2}\, p\alpha} \frac{n^{(3\alpha-1)/2}}{\log n} \longrightarrow \infty. \tag{C.50}$$

Proof. Set $k = 2$ in Eq. (17) and divide by $r_{\text{cr}} = \alpha/[-\log(1-p)]$. Since $\alpha > 1/2$, the exponent $(3\alpha - 1)/2$ is positive, and a positive power of n dominates $\log n$. □

C.8 Logarithmic moments of the overlap kernel

Lemma C.12 (Logarithmic moments of the overlap kernel)**.** Fix $k \geq 2$ and $\alpha \in (0, 1)$, let $q = q_n$ be an integer sequence with $q_n \sim n^\alpha$, and set $\mu = n/q$, and suppose $\omega_n = p_{\text{eff},n}/(1 - p_{\text{eff},n})$ is bounded above and below by positive constants. For every fixed $C_D > 0$, there exist constants $C_k, c > 0$ such that, for all sufficiently large n and all integers $2 \leq t \leq C_D \log n$,

$$\mathbb{E}|\gamma_H|^t \leq \left(\frac{C_k \omega_n \sqrt{t}}{q^k \sqrt{\mu}}\right)^t + 2\omega_n^t e^{-c\mu}. \tag{C.51}$$

Proof. Let $\mathcal{A} = \{|H - \mu| \leq \mu/2\}$. On $\mathcal{A}$,

$$q^k \frac{(H)_k}{(n)_k} = \prod_{i=0}^{k-1} \frac{q(H - i)}{n - i}. \tag{C.52}$$

For each fixed $0 \leq i < k$,

$$\frac{q(H - i)}{n - i} - 1 = \frac{q(H - \mu) - (q - 1)i}{n - i}. \tag{C.53}$$

Because k is fixed and $\mu \to \infty$, every product factor is bounded by 2 in absolute value for all sufficiently large n. Applying

$$\prod_{i=0}^{k-1} u_i - 1 = \sum_{j=0}^{k-1} (u_j - 1) \prod_{i<j} u_i \tag{C.54}$$

therefore gives

$$\left| q^k \frac{(H)_k}{(n)_k} - 1 \right| \leq C_k \frac{|H - \mu| + 1}{\mu} \quad \text{on } \mathcal{A}. \tag{C.55}$$

Since $(q^k - 1)^{-1} \leq 2q^{-k}$ for $q^k \geq 2$,

$$|\gamma_H| \mathbf{1}_{\mathcal{A}} \leq \frac{C_k \omega_n}{q^k} \frac{|H - \mu| + 1}{\mu}. \tag{C.56}$$

For $H \sim \mathrm{Bin}(n, 1/q)$, Bernstein's inequality gives

$$\Pr(|H-\mu| \geq u) \leq 2\exp\left\{-\frac{u^2}{2(\mu+u/3)}\right\}. \tag{C.57}$$

The tail-integral identity

$$\mathbb{E}|H-\mu|^t = t\int_0^\infty u^{t-1}\Pr(|H-\mu| \geq u)\,\mathrm{d}u \tag{C.58}$$

and a split at $u=\mu$ now yield

$$\|H-\mu\|_t \leq C(\sqrt{\mu t}+t). \tag{C.59}$$

Indeed, on $[0,\mu]$ the exponent in Eq. (C.57) is at most $-c_1u^2/\mu$, while on (μ,∞) it is at most $-c_2u$. The two resulting Gamma integrals have the displayed scales. Since $t \leq C_D \log n = o(\mu)$, Eq. (C.56) implies

$$\|\gamma_H \mathbf{1}_{\mathcal{A}}\|_t \leq \frac{C_k\omega_n\sqrt{t}}{q^k\sqrt{\mu}}. \tag{C.60}$$

Finally, Eq. (C.57) gives $\Pr(\mathcal{A}^c) \leq 2e^{-c\mu}$, whereas $|\gamma_H| \leq \omega_n$ everywhere because $0 \leq (H)_k/(n)_k \leq 1$. Splitting the moment over $\mathcal{A}$ and $\mathcal{A}^c$ proves Eq. (C.51). □

C.9 Uniform logarithmic-degree bound

Theorem C.13 (Logarithmic-degree invisibility (restatement of Theorem 3.8))**.** Fix $k \geq 2$ and $1/k < \alpha < 1$. Let $q = q_n$ be an integer sequence with $q_n \sim n^\alpha$, let $p_{\mathrm{eff},n} \to p \in (0,1)$, and assume $p_{\mathrm{eff},n}$ stays away from 0 and 1. Suppose $m = m(n)$ is an integer sequence satisfying $m(n) \leq n^A$ eventually for some fixed $A < \infty$. For any fixed $C_D > 0$ and any integer sequence $D = D(n) \leq C_D \log n$,

$$\Lambda_2 \to 0 \quad \Longrightarrow \quad \|L_{\leq D}\|_2^2 = 1 + o(1). \tag{C.61}$$

Conversely, if $D \geq 2$ and $\Lambda_2 \to \infty$, then

$$\|L_{\leq D}\|_2^2 \to \infty. \tag{C.62}$$

Proof. Set

$$x_n := \frac{m\omega_n}{q^k\sqrt{\mu}} = \frac{m\omega_n}{\sqrt{nq^{2k-1}}}, \qquad T_n := \min\{D, m\}. \tag{C.63}$$

The degree-one term in Eq. (8) is zero. For $2 \leq t \leq T_n$, Lemma C.12 and $\binom{m}{t} \leq (em/t)^t$ give

$$\binom{m}{t}|\mathbb{E}\gamma_H^t| \leq \left(\frac{C_k x_n}{\sqrt{t}}\right)^t + 2\binom{m}{t}\omega_n^t e^{-c\mu}. \tag{C.64}$$

Because $q = o(n)$ and k is fixed, the exact expression Eq. (11) can be written as

$$\Lambda_2 = c_n\omega_n^2\frac{m(m-1)}{nq^{2k-1}}, \qquad c_n \longrightarrow \frac{k^2}{2}. \tag{C.65}$$

For $m \geq 2$, $m^2 \leq 2m(m-1)$. For $m \leq 1$, estimate x_n^2 directly. Thus a constant C independent of the integer sequence $m(n)$ satisfies

$$x_n^2 \leq C\Lambda_2 + \frac{C}{nq^{2k-1}}. \tag{C.66}$$

Hence $\Lambda_2 \to 0$ implies $x_n \to 0$, and the typical part of Eq. (C.64) sums to

$$\sum_{t=2}^{T_n}\left(\frac{C_k x_n}{\sqrt{t}}\right)^t = O(x_n^2) = o(1). \tag{C.67}$$

For the exceptional part, $\omega_n = \Theta(1)$, m is polynomial, and $T_n = O(\log n)$, so

$$\sum_{t=2}^{T_n} 2\binom{m}{t}\omega_n^t e^{-c\mu} \leq 2T_n\exp\{O(T_n\log(m\vee 2)) - c\mu\} \tag{C.68}$$

$$= \exp\{O((\log n)^2) - cn^{1-\alpha}\} = o(1). \tag{C.69}$$

Substitution into Eq. (8) proves Eq. (19). If $\Lambda_2 \to \infty$, orthogonality and Eq. (12) give

$$\|L_{\leq D}\|_2^2 \geq \|L_{\leq 2}\|_2^2 = 1 + \Lambda_2 \to \infty, \tag{C.70}$$

which proves Eq. (20). □

C.10 LOCAL–GLOBAL SEPARATION

Corollary C.14 (Global distinguishability before logarithmic-degree visibility (restatement of Corollary 3.9))**.** Assume the conditions of the classical Model RB threshold, $1/k < \alpha < 1$, $q_n \sim n^\alpha$, $m_n \sim rn \log n$ for a fixed $r > r_{\mathrm{cr}}$, and relation sizes chosen so that $p_{\mathrm{eff},n} \to p \in (0,1)$. Then the satisfiability test distinguishes $\mathbb{P}_1$ from $\mathbb{P}_0$ with error probability tending to zero, while for every fixed $C_D > 0$,

$$\|L_{\le C_D \log n}\|_2^2 = 1 + o(1). \tag{C.71}$$

In particular, every centered degree-$C_D \log n$ statistic with unit variance under $\mathbb{P}_0$ has vanishing planted-versus-null mean separation.

Proof. Every planted instance is satisfiable. For fixed $r > r_{\mathrm{cr}}$, the classical Model RB threshold implies that a null instance is unsatisfiable with probability $1 - o(1)$, so the satisfiability indicator has error $o(1)$ under equal priors.

On the other hand, Eq. (16) and $m = rn \log n$ give

$$\Lambda_2 = \Theta\Big(n^{1-\alpha(2k-1)} (\log n)^2\Big). \tag{C.72}$$

Since $\alpha > 1/k$,

$$1 - \alpha(2k-1) < 1 - \frac{2k-1}{k} = -1 + \frac{1}{k} < 0, \tag{C.73}$$

and therefore $\Lambda_2 \to 0$. Theorem 3.8 proves Eq. (21). Finally, every centered $f \in \mathcal{H}_{\le D}$ with $\mathrm{Var}_0(f) = 1$ satisfies

$$|\mathbb{E}_1 f - \mathbb{E}_0 f| = |\langle L_{\le D} - 1, f \rangle_0| \le \sqrt{\|L_{\le D}\|_2^2 - 1} = o(1). \tag{C.74}$$

□

D EXPERIMENTAL DETAILS

D.1 INSTANCE GENERATION

Instances follow the fixed-cardinality Model RB construction of Section 2 and Appendix B. For each (n, α, p, r), we set

$$q = \mathrm{round}(n^\alpha), \qquad m = \mathrm{round}(rn \log n), \qquad t = \mathrm{round}(pq^2), \tag{D.1}$$

where t is the number of forbidden tuples in each binary relation. Scopes are uniform unordered pairs of distinct variables, sampled independently so that repeated scopes are possible. A null relation forbids a uniform size-t subset of $[q]^2$. A planted instance first draws $\sigma^* \sim \mathrm{Unif}([q]^n)$ and then draws each forbidden set uniformly among the size-t sets that do not contain the planted local tuple. Equivalently, its allowed relation is uniform conditional on containing that tuple, matching the forced-satisfiable generator of Xu et al. (2007). All exact coordinates use $p_{\mathrm{eff}} = t/q^2$.

The classes share (n, q, m, t) and the scope law. They differ only in whether one assignment is admitted by every constraint. We sample fresh instances for each split, and no planted witness is exposed to a model.

D.2 MICROSTRUCTURE GRAPH AND OBSERVATION INTERFACES

The graph $G = (\mathcal{V}, \mathcal{E})$ has one value node (i, x) for every variable $i \in [n]$ and value $x \in [q]$, hence $|\mathcal{V}| = nq$. Values of the same variable form a clique. For a constraint on variables $\{i, j\}$, we add an edge between (i, x) and (j, y) whenever tuple (x, y) is forbidden. Parallel edges created by repeated constraints or repeated forbidden pairs are collapsed. Node features contain normalized degree but no explicit variable identifier, constraint identifier, scope label, or edge multiplicity.

The adjacency may contain structural evidence about the original variable cliques, although their partition is not explicitly represented and its unique recovery after adding cross-variable edges is not established. The relation-level L_2 score below is formed after tuple deduplication but still uses the known variable/scope partition. It therefore tests the loss caused by deduplication, not the entire map from a relation list to the anonymous graph tensor received by a GNN.

D.3 Baseline details

Every learned baseline receives the same anonymous microstructure graph and the same node features, and ends in graph-level pooling followed by an MLP head. The baselines differ in exactly one of three respects: how a neighbourhood is aggregated, how far information travels, and what order of representation is maintained. Holding the input and the head fixed is what lets a difference in accuracy be read as a difference in that one respect.

- **GIN** is the main reference model. Each layer sums neighboring node representations and applies an MLP to the combined self and neighborhood signal. Under the standard injectivity assumptions, this design matches the distinguishing power of 1-WL (Xu et al., 2019). Sum aggregation is particularly relevant here because graph size, degree, and repeated local patterns can affect the representation instead of being normalized away. The main model has three message-passing layers. The depth control uses $2, 3, 4$, and 8 layers to test whether increasing the receptive field, while retaining the same 1-WL-bounded representation order, moves the empirical crossover.
- **GCN** applies a linear transformation after symmetrically degree-normalized aggregation with self-loops (Kipf & Welling, 2017). Its normalization emphasizes smooth neighborhood averages and reduces the direct effect of raw degree relative to GIN. Comparing GCN with GIN tests whether the transition is specific to an injective sum-based update or is also present under a spectral-style smoothing operator.
- **GraphSAGE** uses mean aggregation and combines the resulting neighborhood summary with the representation of the center node (Hamilton et al., 2017). Because the mean discards neighborhood multiplicity unless it changes the empirical feature distribution, GraphSAGE provides a complementary test of whether absolute local counts drive the GIN result.
- **GAT** assigns data-dependent attention weights to adjacent nodes before aggregation (Veličković et al., 2018). This mechanism can prioritize heterogeneous neighbors instead of treating all incident edges uniformly. The comparison tests whether adaptive local weighting can recover planted evidence that fixed sum or mean aggregation fails to emphasize.
- **GIN with a virtual node** introduces a graph-level state that pools information from all nodes and broadcasts it back at every layer. The virtual node greatly shortens communication paths and therefore tests whether limited receptive field or long-range transport explains the hard cells. It remains a message-passing construction, so the shorter communication paths do not increase WL order.
- **PPGN** maintains an order-two tensor indexed by pairs of graph nodes and interleaves feature-wise MLPs with matrix multiplication. The pair-indexed state can represent interactions that standard node-wise message passing cannot capture, making PPGN the principal control for representation order (Maron et al., 2019a). Its dense $|\mathcal{V}| \times |\mathcal{V}|$ state requires $O(|\mathcal{V}|^2)$ memory, which restricts the current runs to $n \leq 25$.

GCN, GraphSAGE, and GAT jointly test whether the main result depends on a particular local aggregation rule. The GIN depth and virtual-node controls test communication range, whereas PPGN changes the order of the maintained representation. These comparisons preserve the conceptual distinction among polynomial interaction degree, message-passing depth, and WL order. In particular, a deeper GIN is neither a higher-order WL model nor automatically a higher-degree detector in the low-degree likelihood-ratio sense.

D.4 Pair score and relation-information ladder

The exact quadratic optimizer on the complete observation is $S_2^* = (L_X)_{=2}$ from Eq. (C.36). The reported diagnostic computes its dominant shared-variable contribution. For a constraint a and incident variable v, let $M_{a,v} \in \mathbb{R}^q$ be the centered marginal-count vector of its forbidden table along the coordinate for v. We evaluate

$$T_{\text{pair}} = \sum_{v \in [n]} \frac{1}{2} \left(\left\| \sum_{a:\, v \in S_a} M_{a,v} \right\|^2 - \sum_{a:\, v \in S_a} \|M_{a,v}\|^2 \right). \tag{D.2}$$

This is the sum of centered marginal inner products over constraint pairs that share a variable. For distinct binary scopes with one-variable overlap, it is the corresponding constructive second-order

term up to a fixed scale. When two constraints repeat the same scope ($J = 2$), the exact optimizer also contains a centered full-table inner product. Eq. (D.2) therefore computes a *shared-variable pair score* rather than the exact finite-n optimizer, which additionally contains the repeated-scope term. Its close empirical calibration to Λ_2 is therefore evidence about the dominant term, rather than a substitute for Theorem 3.4.

The information ladder applies the same score after three successive relation operations:

- L_0: individual constraint tables that retain constraint identity and repeated scopes.
- L_1: tables grouped by scope while retaining multiplicity through summed counts.
- L_2: the binary union of forbidden tuples within each scope, removing multiplicity but retaining the scope/variable partition.

The levels quantify sensitivity of this score to grouping and deduplication. They are not likelihood ratios, despite the mnemonic L_i used in the plots. The empirical squared-SNR ratio in Table E.4 can exceed one because the scores at different levels are separately estimated and rescaled from finite samples. It is consequently a comparison of fitted score scales rather than an information-retention fraction.

D.5 Node features for the encoding ladder

The representation comparison varies only the node features attached to the deduplicated graph $G = \phi(X)$. Every variant uses the identical graph, split, architecture and training protocol, so the comparison isolates what is exposed at each node.

Provenance. Both feature maps are deterministic functions of G alone. Neither reads the hidden assignment, the solution count, a satisfiability oracle, or the label, and neither is fitted on the training split: they are computed per graph before training and are identical for train, validation and test. This is what makes the contrast a statement about how the same observation is read rather than about a change of observation.

Index conventions. Nodes of G are pairs (k, x) with $k \in [n]$ a variable and $x \in [q]$ a value, so $|V| = nq$. Write b for a second variable index. An edge joins (k, x) and (b, y) when some constraint on the scope $\{k, b\}$ forbids the pair (x, y). Edges inside one variable block, which express that a variable takes one value, are present in G but carry no class information and are excluded from the feature computation.

Aligned features (d=3). Let $c_{k,b,x}$ count the edges from node (k, x) into block b, and let $\mathcal{B}_k$ be the set of blocks b with at least one such edge. Centre within each interacting block pair,

$$M_{k,b,x} = \Big(c_{k,b,x} - \tfrac{1}{q}\textstyle\sum_{x'} c_{k,b,x'}\Big)\mathbf{1}\{b \in \mathcal{B}_k\}, \tag{D.3}$$

and form the per-variable power sums

$$S_{k,x} = \sum_b M_{k,b,x}, \qquad Q_k = \sum_{b,x} M^2_{k,b,x}, \qquad T_k = \tfrac{1}{2}\Big(\textstyle\sum_x S^2_{k,x} - Q_k\Big). \tag{D.4}$$

Node (k, x) receives $(S_{k,x},\ S^2_{k,x},\ T_k)$. The centring in Eq. (D.3) is what removes the null component, since exact quiet planting makes the uncentred block marginal label-independent. The term T_k is the per-variable summand of the shared-variable quadratic score, so the model is handed the local evidence the theory identifies and must still learn how to combine and pool it across variables.

Generic features (d=3). The control attaches (triangles, wedges, four-cycles) at each node, computed on the same G: the triangle count $\frac{1}{2}(A^2 \odot A)\mathbf{1}$, the wedge count $\binom{\deg}{2}$, and the four-cycle count obtained from common-neighbour multiplicities. These are the substructure counts used by substructure-enriched message passing (Bouritsas et al., 2023; Chen et al., 2020). They have the same dimension and the same computational interface as the aligned features and differ only in being unrelated to the generating mechanism.

Scaling. Both feature sets are standardised per graph to zero mean and unit variance across nodes, so that feature magnitude cannot substitute for the signal, and so that the two variants enter the network on the same scale.

Base feature and totals. Every rung of the representation comparison starts from the same one-dimensional base feature, the node degree of G divided by the largest degree in that graph. The *anonymous* rung uses it alone, $d{=}1$; the aligned and generic rungs append their three standardised features to it, $d{=}4$ each. The base feature is itself informative, since the degree of (k, x) is the raw number of constraints that forbid value x for variable k; what the aligned features add on top of it is the split by neighbouring variable and the centring within each block pair, which is exactly the part quiet planting makes necessary. The *factor-graph* rung replaces G by the bipartite variable–constraint incidence graph, whose two node types are marked by a two-dimensional indicator, $d{=}2$; it exposes the scope partition that anonymisation removes, without exposing the relation tables.

D.6 Negative and shallow controls

The controls in this subsection address a basic attribution question: could the reported AUC arise without learning consistency induced by one shared planted assignment? We test three increasingly structured alternatives. Label shuffling removes every population-level association between an instance and its class label. Independent local planting preserves the one-constraint generation rule but removes cross-constraint consistency. Shallow graph summaries retain the true labels, but restrict the classifier to coarse size and collision statistics. The three controls therefore probe evaluation noise, marginal-generation artifacts, and trivial graph-encoding artifacts, respectively.

Label shuffle. We randomly permute labels while leaving the instances and evaluation pipeline unchanged. Since the shuffled label is independent of every graph statistic, its population AUC is 0.5. The largest observed AUC over the complete grid is 0.546. Because the maximum is taken over many cells, it calibrates the finite-sample excursions that occur when multiple configurations are inspected.

Independent local planting. For every constraint independently, we draw a uniform local tuple and then sample a relation conditional on allowing that tuple. This procedure matches the local conditioning step used by planted Model RB, but different constraints no longer inherit their tuples from one globally consistent assignment. Lemma B.4 proves that marginalizing these independent tuples recovers exactly $\mathbb{P}_0$. Thus no detector, regardless of its capacity, can achieve population AUC above 0.5 against the null class. The largest finite-sample AUC across all cells is 0.551. Agreement with the label shuffle control verifies both the quiet one-constraint marginal and the fact that the useful pair signal depends on consistency across constraints rather than on the local conditioning operation alone.

Shallow graph summaries. The microstructure encoding can create class-dependent graph summaries through finite-size collisions and edge deduplication, even when no subtle relational pattern is extracted. We therefore evaluate the edge count, degree variance, wedge count, and a cross-validated logistic classifier on the degree histogram. The wedge count is the sum of the number of unordered neighbor pairs at each node, so it measures second-order degree concentration without using node or constraint identities. These statistics ask whether a label can be predicted from coarse graph density alone. They are diagnostic controls rather than expressive graph-learning baselines: strong performance would reveal a trivial encoding shortcut, whereas near-chance performance would rule out these particular summaries without proving that all anonymous-graph shortcuts are absent. Table E.3 shows that such artifacts are substantial in a few dense, small instances but weaken rapidly as n increases.

The controls establish a hierarchy of conclusions. The label shuffle verifies the evaluation floor. Independent local planting shows that the signal is genuinely cross-constraint. Shallow summaries determine when coarse properties of the encoded graph suffice. None of these controls is a replacement for exact 1-WL or a graph-measurable pair detector, both of which remain necessary for a complete expressivity attribution.

D.7 Training and evaluation protocol

We train every model with hidden width 64, dropout 0.5, Adam at learning rate 10^{-3} with weight decay 5×10^{-4}, a cosine learning-rate schedule, and 200 epochs (Kingma & Ba, 2015). Datasets contain 2000 graphs per class for $n \in \{15, 25\}$, 1500 per class for $n = 40$, and 1000 per class for $n = 60$. We use a stratified $60/20/20$ train/validation/test split that is disjoint at the instance level. Checkpoints are selected by validation accuracy and evaluated once on the test split. Each neural configuration is run with three model seeds on the same data split, and tables report mean and standard deviation. Departures from these defaults are stated where they occur.

The detector and control suite uses 500 instances per class per cell, and the encoding ladder uses 400 per class. AUC is computed from held-out scores. Whenever an accuracy is reported for an uncalibrated score, its sign and threshold are selected on one half of the detector sample and evaluated on the other half. Empirical crossings are obtained by linear interpolation between adjacent density values and are reported as $> r_{\max}$ when the requested level is not reached. With three neural seeds, standard deviations are descriptive.

D.8 Experimental setting

Hard and learnable cells. Several controls report averages restricted to the cells where the question is live. We call a cell (n, r) of a grid *hard* when the reference GIN on the anonymous graph attains less than 60 AUC there, and *learnable* when it exceeds 75. The thresholds are a reporting convention rather than an estimated boundary, and the grouping is post hoc: it is defined by the reference model's own outcome, so it measures whether an intervention changes that outcome. Each table states its reference, since the backbone control uses the main-grid GIN while the encoding and architecture ladders use the anonymous-graph GIN trained in the same sweep.

Compute budget. One (grid, size, model) combination is one stage of a sweep, covering nine densities and three seeds, and each stage is allotted 72 hours on one GPU. A stage that does not finish is marked OOT and its cells are reported as such rather than averaged over the subset that did finish. Which stages are OOT follows from instance size and architecture cost, fixed before the runs, so the omission does not select on the outcome. The $(\alpha, p) = (0.8, 0.3)$ grid at $n = 40$ is representative: its four stages take 55.4 hours for the four encodings under GIN, 9.3 for the virtual node, 16.8 for the graph-statistics MLP, and more than 15.7 for the node-marking Subgraph GNN. The Subgraph GNN dominates because node marking runs one forward pass per root, eight per instance, over a graph that already has 760 nodes and 37,570 edges at this cell.

The encoding and architecture ladders are therefore reported at $n \in \{15, 25, 40\}$ for $(\alpha, p) \in \{(0.6, 0.5), (0.6, 0.3), (0.8, 0.3)\}$ and at $n \in \{15, 25\}$ for $(0.8, 0.5)$. The main GIN sweep, the backbone sweep and the explicit statistics are cheaper and cover all four sizes.

E Extended Experimental Results

In this section we extend the experiments summarised in the main text. We first resolve the empirical crossing and the calibration residual in more detail, then report the controls that diagnose the extraction gap, and finally give the full encoding and architecture ladders together with the backbone sweep and the densities that lie beyond the swept range.

E.1 Sensitivity of the empirical crossing

The empirical crossing $r_{\mathrm{GIN}}^{(\theta)}$ is defined by linear interpolation of the measured curve to a target level θ, so it exists only when the sweep actually reaches θ. Entries written $> r_{\max}$ are the cells where it does not: at the densest swept point the curve is still below the requested level, so the crossing lies outside the sampled grid and is right-censored rather than missing. Those cells are discussed in Appendix E.8.

To verify that the size-dependent drift is not an artifact of choosing one metric or one operating point, we recompute the empirical crossing using accuracy and AUC at three target levels. Table E.1 reports the resulting sensitivity analysis.

Table E.1: Sensitivity of the empirical GIN crossing to the evaluation metric and selected performance level. The ordering across n is unchanged within the sampled grid. $> r_{\max}$ means that the requested level is not reached.

(α, p)	Metric	Level	n			
			15	25	40	60
$(0.6, 0.3)$	Acc	65	2.11	2.82	3.71	6.27
		70	2.42	3.51	4.14	6.98
		75	2.80	4.29	5.19	8.12
	AUC	65	1.86	2.32	3.23	5.78
		70	2.09	2.68	3.59	6.31
		75	2.32	3.17	3.97	6.93
$(0.8, 0.3)$	Acc	65	6.10	8.76	$> r_{\max}$	$> r_{\max}$
		70	6.98	9.89	$> r_{\max}$	$> r_{\max}$
		75	7.68	11.03	$> r_{\max}$	$> r_{\max}$
	AUC	65	5.27	7.75	$> r_{\max}$	$> r_{\max}$
		70	5.91	8.52	$> r_{\max}$	$> r_{\max}$
		75	6.61	9.43	$> r_{\max}$	$> r_{\max}$
$(0.6, 0.5)$	Acc	65	1.20	1.94	2.69	4.44
		70	1.44	2.26	3.09	$> r_{\max}$
		75	1.65	2.61	3.49	$> r_{\max}$
	AUC	65	0.93	1.71	2.03	3.34
		70	1.15	1.89	2.68	4.21
		75	1.36	2.17	2.95	$> r_{\max}$
$(0.8, 0.5)$	Acc	65	4.01	6.09	$> r_{\max}$	$> r_{\max}$
		70	4.55	6.87	$> r_{\max}$	$> r_{\max}$
		75	5.14	$> r_{\max}$	$> r_{\max}$	$> r_{\max}$
	AUC	65	3.55	5.51	10.15	$> r_{\max}$
		70	3.96	6.07	$> r_{\max}$	$> r_{\max}$
		75	4.37	6.62	$> r_{\max}$	$> r_{\max}$

Within every parameter grid and every reported metric-level combination, the crossing moves to larger density as n increases or leaves the sampled range. For example, in the $(0.6, 0.5)$ grid the AUC-0.70 crossing increases from 1.15 to 4.21, while in the $(0.8, 0.5)$ grid it is beyond the sweep for $n \geq 40$. The numerical location changes with the metric and target level, as expected for a finite-size crossover, but its drift with problem size does not.

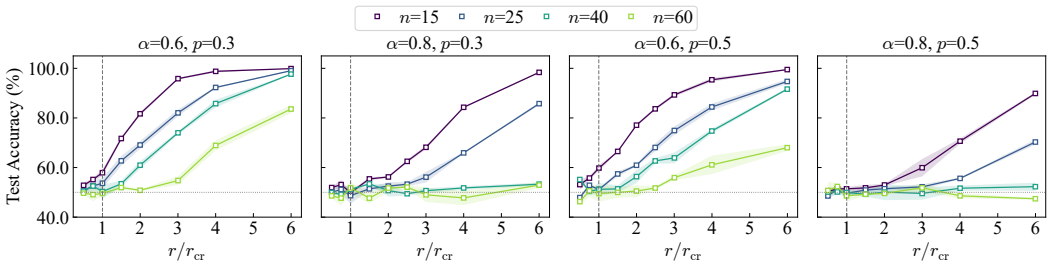


Figure E.1: GIN test accuracy against density normalized by the satisfiability threshold. Curves show means over three seeds, with one standard deviation.

Figure E.1 gives the accuracy counterpart of the main-text AUC curves. The same rightward movement with n appears under the fixed decision threshold, confirming that the crossover is not induced by the threshold-free evaluation metric.

Table E.2: Hard cells grouped by Λ_2. The Gaussian reference and the implemented pair score describe relation-level evidence. The remaining columns evaluate the same GIN with anonymous, generic structural, or relation-aligned features. The generic and aligned inputs have matched feature dimension, which does not make their information content or effective capacity identical. Only sizes whose nine-density sweep completed are included. Values are test AUC in percent.

Λ_2	Cells	Available Evidence		GIN on the Same Graphs		
		$\Phi(\sqrt{\Lambda_2/2})$	Pair Score	Anonymous	+ Generic	+ Aligned
< 1	35	65.78	64.38	52.38	51.37	63.05
$[1, 3)$	14	82.65	79.98	55.05	53.18	92.78
≥ 3	4	94.51	91.86	55.08	52.04	99.81

E.2 Calibration residual against the Gaussian reference

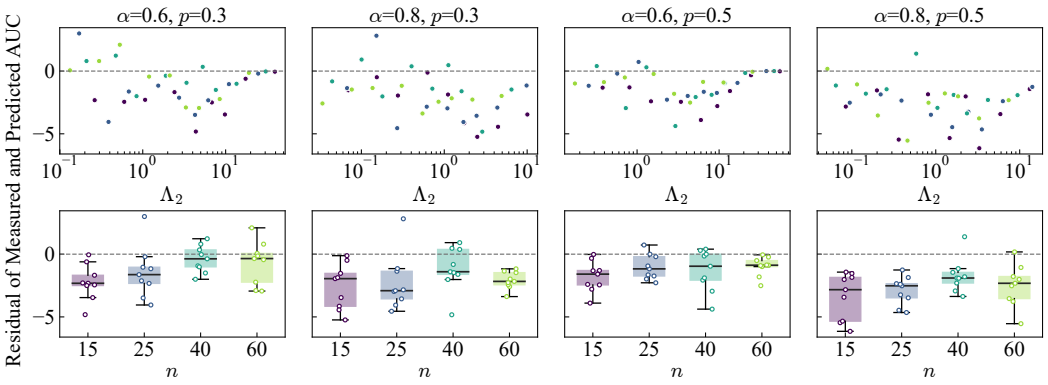


Figure E.2: Measured minus predicted AUC for the implemented shared-variable pair score, over all 36 cells of each grid, with the prediction taken from the Gaussian reference $\Phi(\sqrt{\Lambda_2/2})$. Columns are the four grids and both rows show the same values, plotted signed against the dashed zero line. The upper row resolves them along Λ_2 and the lower row groups them by instance size. Colours encode n in both rows.

The two rows test two different ways the calibration could fail, and neither failure appears.

The upper row asks whether the residual is systematic along Λ_2 itself. If it drifted monotonically, Λ_2 would not be a sufficient coordinate and a second variable would be needed. It does not. Binned by Λ_2, the mean residual is -1.10 below 0.3, then -1.64, -2.29 and -2.57 through the intermediate bins, and -0.71 above 10. The deficit is therefore deepest in the middle and narrows at both ends, which is expected rather than informative there: for very small Λ_2 both quantities sit near 0.5 and for very large Λ_2 both approach 1, so agreement is forced by the range of AUC in either tail. The intermediate regime is where the comparison has room to fail, and the residual stays within about 2.5 points.

The lower row asks whether the residual grows with instance size. This is the more consequential test, because a residual that grew with n would mean the collapse in Figure 4 is an artifact of small instances. Again it does not: the mean residual is -2.49 at $n = 15$, -1.92 at $n = 25$, -1.14 at $n = 40$ and -1.55 at $n = 60$. It narrows rather than widens over the sizes we can reach.

The largest discrepancy over all four grids is 6.15 AUC points. Plotting the deviation with its sign also shows that it is not symmetric: the implemented score falls below the Gaussian reference in 126 of the 144 cells, and the grid means are negative throughout, from -1.19 to -2.61 points, with the largest deficits at $n = 15$. The one-sided deficit is consistent with the fact that the implemented statistic is a fixed shared-variable contraction rather than the exact finite-n optimizer, although this direction does not follow from Theorem 3.4, which controls SNR rather than AUC. The agreement

Table E.3: Largest AUC attained by any shallow graph summary over the full density sweep. Deduplication artifacts are strongest for dense, small instances and weaken with increasing n.

(α, p)	Metric	Best Shallow Statistic Over the Whole r Sweep			
		$n = 15$	$n = 25$	$n = 40$	$n = 60$
$(0.6, 0.3)$	AUC	98.57	79.41	66.99	62.49
$(0.8, 0.3)$	AUC	77.11	59.33	56.14	53.80
$(0.6, 0.5)$	AUC	82.20	64.64	57.07	55.64
$(0.8, 0.5)$	AUC	61.46	56.22	55.98	53.48

supports the use of Λ_2 as an empirical coordinate for the implemented shared-variable score. It does not turn the Gaussian reference into a theorem and does not make Λ_2 a bound on a graph classifier.

To separate weak finite-sample pair evidence from evidence that is already visible to the relation-level detector, Table E.2 groups hard cells by Λ_2. The cut points are descriptive conventions. In particular, $\Lambda_2 < 1$ is not the asymptotic premise $\Lambda_2 \to 0$ of Theorem 3.8.

E.3 Shallow graph summaries

This control asks whether the task can be solved from coarse properties of the encoded graph alone, without relational reasoning. If simple counts sufficed, the crossover could not be attributed to cross-constraint evidence. The answer depends on the regime.

Small-n density artifacts. The best edge/degree/wedge baseline reaches AUC 0.986 for a dense $n = 15$ grid, but its maximum falls to 0.625 by $n = 60$ and to at most 0.556 in two of four $n = 60$ grids (Table E.3). Thus simple encoding artifacts affect some small instances but do not explain the size-dependent crossover.

The small-instance result is an important qualification. Some dense cells can be classified using finite collision or graph-density artifacts and therefore cannot be attributed uniquely to relational reasoning. At $n = 60$, however, the maxima range from 53.48 to 62.49 across the four grids. The measured coarse summaries consequently do not explain the persistent large-instance gap between the relation-aware score and GIN.

E.4 Controls on the extraction gap

An earlier crossover is informative only when the intervention identifies what changed. We therefore separate four mechanisms. The relation-information ladder tests whether grouping and tuple deduplication remove the signal. The choice of message-passing backbone tests the local aggregation rule. Greater depth and a virtual node test communication range while retaining an order-one message-passing representation. PPGN tests a pair-indexed representation, while data scaling and train–test comparisons probe estimation and optimization. Appendix D specifies each interface and training protocol.

Information interface and the extraction gap. To isolate grouping and deduplication from the final anonymous graph interface, we evaluate the same shared-variable score at three relation-information levels. L_0 retains individual relation tables, L_1 groups tables by scope while retaining multiplicity, and L_2 forms the binary union of forbidden tuples for each scope. Table E.4 and Figure E.3 report the comparison.

Table E.4: Relation-information ladder. The AUC columns compare the shared-variable score before and after grouping and tuple deduplication. "SNR^2 ratio" is a finite-sample diagnostic rather than an information-retention fraction. "Control" is the larger maximum from independent local planting and label shuffling.

(α, p)	Cells	Mean $\pm$ Std of $\lvert A_{L_0} - A_{L_2} \rvert$	Max	Median Retained	Control (Max)
$(0.6, 0.3)$	36	0.98 $\pm$0.84	3.18	1.16	54.65
$(0.8, 0.3)$	36	2.04 $\pm$1.92	7.19	1.20	54.06
$(0.6, 0.5)$	36	0.59 $\pm$0.48	1.51	1.06	55.44
$(0.8, 0.5)$	36	1.59 $\pm$1.32	5.48	1.14	54.82

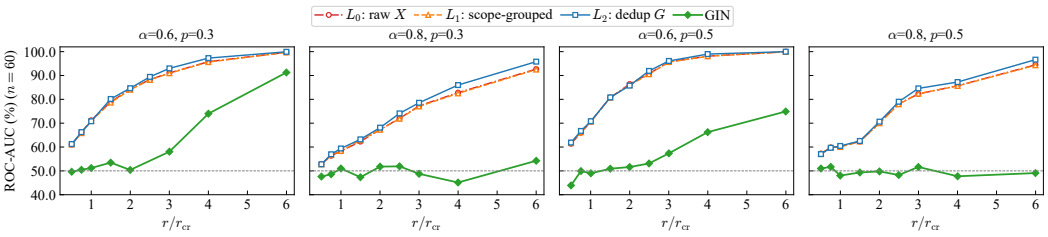


Figure E.3: Relation-information ladder at $n = 60$. The shared-variable score is evaluated on individual constraint tables (L_0), scope-grouped tables that retain multiplicity (L_1), and binary unions that remove multiplicity (L_2). The curves remain close. The L_2 score nevertheless uses variable and scope metadata that are absent from the anonymous graph supplied to GIN.

Across the four grids, the mean absolute L_0-to-L_2 AUC change lies between 0.59 and 2.04 points, with a largest observed change of 7.19 points. Thus the removal of repeated relation tuples does not account for most of the larger pair-score-to-GIN discrepancy. The median empirical SNR-squared ratios slightly exceed one in some grids because the finite-sample scores are estimated and rescaled separately, not because deduplication creates information. Since L_2 still uses the variable and scope partition, this experiment addresses deduplication but does not close the broader question of what is lost when the instance is presented as an anonymous graph.

We call the L_2-score-to-GIN residual an *empirical extraction gap*. It is not a computational–statistical gap because the two procedures receive different observations. Figure 5 in the main text measures this residual throughout each density sweep, and Figure E.4 places its trend beside the global satisfiability reference, the unit-SNR contour, and the two empirical crossings.

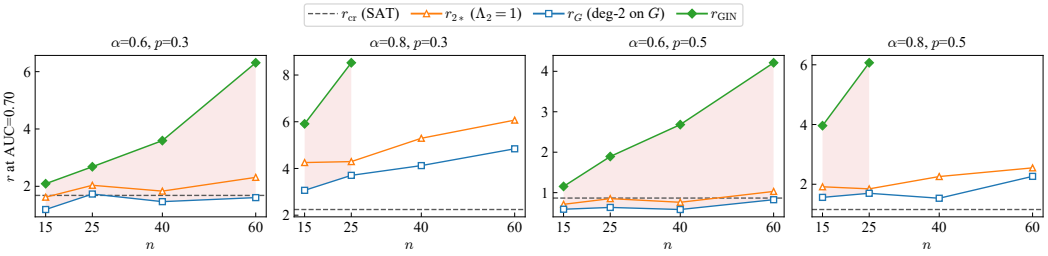


Figure E.4: Four reference contours in the (n, r) plane. The satisfiability threshold r_{cr} and the unit-SNR contour $r_{2,*}$, where $\Lambda_2 = 1$, are analytic, while the AUC-0.70 crossing r_G of the deduplicated relation-level pair score and the empirical GIN crossing $r_{\mathrm{GIN}}^{(0.70)}$ are measured. Because the four contours use different nominal levels, the comparison concerns their scale and trend rather than equality. Shading denotes the observed relation-to-GIN extraction gap.

Table E.5: Backbone control on the $(\alpha, p) = (0.6, 0.3)$ grid. Cells are grouped post hoc by GIN AUC. The comparison tests whether the observed hard regime is specific to one neighborhood aggregation rule.

Architecture	Hard Regime ($A_{\text{GIN}} < 60$)		Learnable ($A_{\text{GIN}} > 75$)	
	Acc	AUC	Acc	AUC
GIN	51.99 $\pm$1.98	53.23 $\pm$2.39	85.45 $\pm$10.93	91.10 $\pm$8.50
GAT	50.35 $\pm$1.42	50.59 $\pm$2.15	74.13 $\pm$17.32	78.11 $\pm$17.50
GCN	50.60 $\pm$0.95	50.51 $\pm$2.25	61.56 $\pm$8.41	66.07 $\pm$10.42
GraphSAGE	53.18 $\pm$2.99	55.23 $\pm$4.66	87.35 $\pm$10.47	92.77 $\pm$7.84

The maximum gap increases from 15.86 to 38.81 AUC points between $n = 15$ and $n = 60$ for $(\alpha, p) = (0.6, 0.5)$, and from 25.72 to 47.57 points for $(0.8, 0.5)$. The mean gap also grows strongly in the two $\alpha = 0.6$ grids. The trend is not exactly monotone: for $(0.8, 0.3)$ the maximum decreases from 43.81 at $n = 40$ to 41.57 at $n = 60$.

Resolving the gap along the sweep rather than summarising it reveals a common shape in the grids whose turnover is observed: the gap starts small, rises to a maximum, and then decays as GIN reaches the relation-level score. Two things generally change with size. The peak grows from 16 to 39 points across n for $(0.6, 0.5)$ and from 15 to 35 for $(0.6, 0.3)$. The peak also moves to higher density, from $1.5\, r_{\text{cr}}$ to $2.5\, r_{\text{cr}}$ for $(0.6, 0.5)$ and from $0.75\, r_{\text{cr}}$ to $3\, r_{\text{cr}}$ for $(0.6, 0.3)$, with a single inversion between $n = 15$ and $n = 25$ in the former. The decay also slows: at the densest swept point the gap has closed to under 3 points for $n \leq 40$ in $(0.6, 0.5)$ but still stands at 25 points for $n = 60$. The band in which relation-level evidence is present and GIN does not use it therefore both widens and shifts outward with instance size, which is the same statement the crossing contours of Figure E.4 make in a different coordinate.

The two $\alpha = 0.8$ grids qualify this picture in an informative way. There the $n = 40$ and $n = 60$ curves have not turned over by the end of the sweep: the gap is still at its maximum, 45 and 48 points for $p = 0.5$, at $6\, r_{\text{cr}}$. These are precisely the four cells whose GIN crossing lies beyond the swept density range (Appendix E.8), so the two observations are the same fact seen from two directions, and the reported maxima for those cells are lower bounds rather than located peaks. The appropriate claim is therefore a general increase with size, accompanied by finite-sample and grid-dependent deviations.

Local aggregation rules. To determine whether the hard regime is specific to GIN's sum aggregation, we compare GIN with GCN, GraphSAGE, and GAT after grouping cells in the $(0.6, 0.3)$ grid by the reference GIN AUC. The grouping is post hoc, so Table E.5 is a robustness diagnostic rather than an independent partition of the problem space.

In the hard group, the mean AUCs of GIN, GAT, GCN, and GraphSAGE lie between 50.51 and 55.23. In the learnable group they rise to between 66.07 and 92.77, with substantial variance. The shared near-chance behavior in hard cells shows that the phenomenon is not unique to GIN's aggregation rule. The spread in learnable cells also shows that architecture and optimization affect how efficiently available evidence is extracted. Since all four models are message-passing architectures and the grouping uses GIN, this result is not an architecture-independent lower bound.

Data scale and generalization. Finite data is the remaining ordinary explanation for the gap, and testing it requires more than showing that additional examples fail to help: a sweep in which nothing moves is equally consistent with a probe too blunt to detect any data effect. We therefore hold the model family and the problem cells fixed, increase the training set eightfold from 500 to 4000 instances per class, and include a cell in which the signal is already within the model's reach as a positive control. If the control responds while the hard cells do not, the sweep is demonstrably sensitive and the absence of a response elsewhere carries information.

The four cells span the range of behaviour we want to separate. At n=60, r=2.16 the anonymous-graph GIN sits at chance; at n=40, r=1.73 it is near chance; at n=40, r=2.6 it is partially learnable; and at n=40, r=3.46 it already classifies well, which makes that cell the positive control. A sweep in which nothing moves anywhere would be uninformative, because it is equally consistent with a probe too blunt to detect any data effect; the control is what rules that out.

Table E.6: Effect of training-set size on GIN at $(\alpha, p) = (0.6, 0.5)$, in percent (mean $\pm$ standard deviation over three seeds). "Relation score" is the AUC of the deduplicated relation-level pair statistic on the same underlying instances, and "shortfall" is its distance from GIN at 4000 instances; both are reported for AUC only, since the pair statistic has no intrinsic decision threshold. The relation score is an achievable value on the same observation, not a certified ceiling for graph-level classifiers.

Cell	Metric	Training Graphs per Class				500→4000	Relation Score	Shortfall
		500	1000	2000	4000			
n=60, r=2.16	Acc	51.00 ±1.32	54.33 ±1.89	51.58 ±2.58	54.17 ±0.69	+3.17	–	–
	AUC	48.87 ±1.33	55.22 ±0.13	51.70 ±2.13	55.05 ±0.63	+6.19	91.92	36.87
n=40, r=1.73	Acc	50.67 ±4.22	56.21 ±2.28	55.62 ±1.10	59.19 ±3.21	+8.52	–	–
	AUC	51.05 ±4.27	59.65 ±2.09	58.91 ±1.40	63.22 ±3.93	+12.17	92.82	29.60
n=40, r=2.6	Acc	59.75 ±2.62	61.25 ±2.83	66.96 ±2.19	75.00 ±0.74	+15.25	–	–
	AUC	64.45 ±3.47	66.62 ±3.44	73.44 ±1.47	81.94 ±0.60	+17.49	98.57	16.63
n=40, r=3.46	Acc	58.67 ±3.82	67.33 ±2.52	77.71 ±1.20	82.40 ±0.96	+23.73	–	–
	AUC	64.40 ±3.99	74.19 ±1.98	86.33 ±0.27	90.73 ±0.55	+26.33	99.84	9.11

Three readings follow from Table E.6. First, the four cells are comparable in how much evidence their graphs contain: the relation score lies between 91.92 and 99.84 throughout, so differences in outcome cannot be attributed to one cell simply carrying less signal. Second, the response to data is strongly ordered. Eightfold more data is worth +26.33 AUC points in the learnable control at r=3.46 and +17.49 in the partially learnable cell at r=2.6, but only +12.17 and +6.19 in the two hard cells. Third, the residual distance to the relation score follows the same ordering and remains large exactly where the response is weak: the shortfall falls monotonically from 36.87 to 29.60, 16.63 and 9.11 points across the four cells.

The control is what licenses the interpretation. Because eight times more data nearly closes the gap in the control cell, the sweep can detect a data effect of the relevant magnitude; the near absence of such an effect in the hard cells is therefore a measurement rather than a null result. The accompanying accuracy columns show the same pattern, so the ordering is not an artifact of the threshold-free metric.

We conclude that additional data helps where the model can already express the discriminative function and does not where it cannot. Sample size contributes to the observed crossover, which is one reason we describe it as a finite-size crossover rather than a sharp threshold, but it does not account for the residual gap within the tested range. This complements the training-accuracy evidence in Figures 6 and E.5: the hard cells are not cases in which the network fails to fit, nor cases in which it has too few examples, but cases in which what it fits does not transfer.

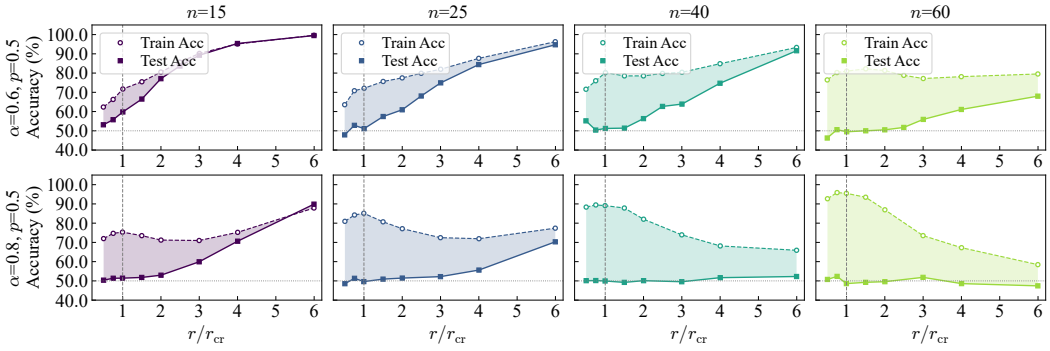


Figure E.5: GIN training accuracy (dashed, open markers) and test accuracy (solid, filled markers) across density for the two $p = 0.5$ grids. Shaded regions show the train–test gap; the corresponding $p = 0.3$ grids appear in Figure 6.

Optimization and memorization. Figures 6 and E.5 jointly resolve training and test accuracy across every parameter grid and instance size. The harder cells exhibit a visible train–test gap. The network can fit more training labels than it predicts on held-out instances, so inability to reduce training error is not the sole cause of chance-level test performance. This observation does not prove that the optimizer found the best generalizing GIN solution and does not distinguish overfitting from a mismatch between the anonymous graph representation and the target signal.

We call this residual an *empirical extraction gap*. It is not a computational–statistical gap, because the two procedures receive different observations. Section 4.3 takes the question further by varying what the graph encodes and what the architecture can separate as two independent axes.

E.5 Communication range and representation order

Communication range and representation order. Finally, we vary GIN depth, introduce a virtual node, and evaluate PPGN at three densities. Greater depth and the virtual node change information transport while retaining node-wise message passing. PPGN instead maintains a pair-indexed state. Figure E.6 reports the comparison.

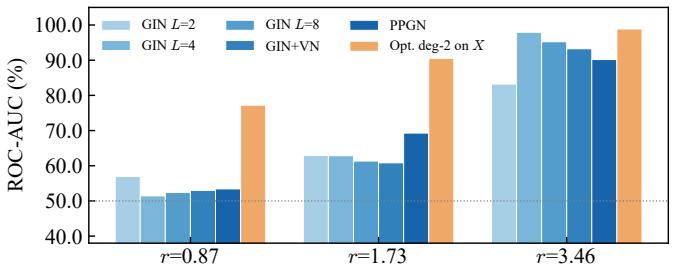


Figure E.6: Architecture comparison at $n = 25$ and $(\alpha, p) = (0.6, 0.5)$. Depth and a virtual node do not consistently improve the lower-density cells. The large PPGN error bars prevent a definitive claim about higher representation order.

At the lowest density, the shared-variable score reaches 77.22 AUC while all learned models lie between 51.40 and 56.95. At the intermediate density, the GIN variants and virtual-node model remain between 60.82 and 65.71 AUC. PPGN reaches a higher mean of 69.28, but its standard deviation is 18.39 points. Communication range alone is therefore insufficient for the evaluated architectures and cells. The PPGN variance, limited size range, and three-seed evaluation prevent a conclusion about whether higher-order learning moves the crossover earlier. This remains a central empirical question rather than a claim established by the current results.

E.6 Disentangling representation and architecture

To determine which interventions recover the extraction gap, we vary the representation and learner in two controlled comparisons while holding the instances, splits, and training protocol fixed. Both comparisons use the hard cells, where recovery beyond the anonymous-graph GIN is informative.

Information exposed by the representation. We first fix GIN and vary its input. The comparison progresses from the anonymous deduplicated graph to generic structural features, a factor graph that exposes variable–constraint incidence, and features aligned with shared-variable relations. Figure 7 shows that generic features provide little improvement, the factor graph gives an intermediate gain, and relation-aligned features recover the most signal. This ordering isolates the benefit of exposing the organization of local constraints rather than merely adding feature dimensions. Table E.7 reports every completed representation comparison; incomplete sweeps are marked OOT rather than averaged over a truncated density range.

Table E.7: Representation comparison at fixed GIN architecture. The columns compare the anonymous graph, generic structural features, explicit variable–constraint incidence, and shared-variable-aligned features. Values are test AUC in percent, reported as mean ± standard deviation over the hard cells of each grid and three seeds. "Hard r" is the number of included densities; OOT marks an incomplete stage that exceeded the 72-hour budget.

(α, p)	n	Hard r	Anonymous (d=1)	Generic feat. (d=4)	Factor graph (d=2)	Aligned feat. (d=4)
$(0.6, 0.3)$	15	2	55.52 ±2.23	55.71 ±0.67	60.54 ±4.03	65.35 ±4.48
	25	3	54.27 ±2.23	53.82 ±3.61	60.12 ±6.18	64.37 ±8.33
	40	3	52.47 ±2.09	50.54 ±3.21	59.47 ±5.91	66.47 ±8.38
	60	OOT	OOT	OOT	OOT	OOT
$(0.8, 0.3)$	15	4	52.84 ±3.15	50.96 ±1.78	57.05 ±6.47	61.82 ±10.52
	25	7	54.05 ±3.88	52.26 ±2.58	60.77 ±9.76	75.22 ±19.59
	40	9	52.00 ±3.20	51.21 ±2.73	57.98 ±11.01	77.37 ±19.11
	60	OOT	OOT	OOT	OOT	OOT
$(0.6, 0.5)$	15	2	55.84 ±2.23	55.26 ±3.65	63.75 ±3.80	67.46 ±6.49
	25	4	53.44 ±4.71	52.68 ±2.52	65.04 ±9.38	73.90 ±11.63
	40	5	54.51 ±1.75	51.64 ±2.17	68.72 ±13.16	81.49 ±16.00
	60	OOT	OOT	OOT	OOT	OOT
$(0.8, 0.5)$	15	6	53.78 ±3.84	52.46 ±3.02	61.39 ±10.88	72.91 ±17.88
	25	8	51.82 ±2.75	50.18 ±2.12	65.28 ±14.11	79.48 ±19.77
	40	OOT	OOT	OOT	OOT	OOT
	60	OOT	OOT	OOT	OOT	OOT

Extraction at a fixed representation. We next hold the anonymous graph fixed and vary the learner, comparing a statistics MLP, GIN, GIN with a virtual node, and a node-marking Subgraph GNN. As summarized in the lower row of Figure 7, these changes produce smaller and less stable improvements, and the Subgraph GNN advantage decreases with problem size. Table E.8 reports every completed architecture comparison. The result does not preclude a different architecture from recovering the missing contraction, but it shows that the evaluated generic increases in communication or expressivity are less effective than exposing the relations that carry the signal.

Table E.8: Architecture comparison at fixed anonymous graph representation. The methods progress from a graph-statistics MLP to a 1-WL-bounded GIN, GIN with a virtual node, and a node-marking Subgraph GNN. Values are test AUC in percent, reported as mean ± standard deviation over the hard cells of each grid and three seeds. Conventions follow Table E.7; OOT marks an incomplete stage that exceeded the 72-hour budget (Appendix D.8).

(α, p)	n	Hard r	Statistics MLP	GIN (1-WL)	GIN+VN	Subgraph GNN
$(0.6, 0.3)$	15	2	54.22 ±4.41	55.52 ±2.23	56.13 ±1.37	61.04 ±1.57
	25	3	52.28 ±0.84	54.27 ±2.23	51.72 ±1.20	56.20 ±3.99
	40	3	49.56 ±2.39	52.47 ±2.09	50.61 ±1.49	50.85 ±2.79
	60	OOT	OOT	OOT	OOT	OOT
$(0.8, 0.3)$	15	4	49.59 ±2.96	52.84 ±3.15	49.18 ±1.83	53.07 ±1.81
	25	7	52.07 ±3.08	54.05 ±3.88	50.83 ±1.49	53.18 ±4.73
	40	9	51.33 ±2.19	52.00 ±3.20	51.20 ±1.90	50.38 ±2.64
	60	OOT	OOT	OOT	OOT	OOT
$(0.6, 0.5)$	15	2	50.33 ±1.21	55.84 ±2.23	50.65 ±1.10	56.78 ±1.77
	25	4	51.01 ±2.56	53.44 ±4.71	50.93 ±2.60	54.78 ±4.74
	40	5	50.28 ±3.30	54.51 ±1.75	51.29 ±3.66	53.59 ±3.84
	60	OOT	OOT	OOT	OOT	OOT
$(0.8, 0.5)$	15	6	52.40 ±1.94	53.78 ±3.84	53.40 ±4.11	54.90 ±5.13
	25	8	50.59 ±2.93	51.82 ±2.75	51.15 ±2.63	52.20 ±2.79
	40	OOT	OOT	OOT	OOT	OOT
	60	OOT	OOT	OOT	OOT	OOT

Finally, the explicit detectors in Table E.9 separate alignment from polynomial order without learning. At matched order, the aligned second-order statistic substantially outperforms generic motif counts, whereas raising an aligned statistic from second to third order does not improve the result.

Table E.9: Explicit statistics on 500 instances per class, AUC in percent, averaged over the density sweep at each size. $|S_1|$ is the folded sum of the centred relation marginals. Folding is an even operation, so $|S_1|$ is not a degree-one polynomial in the instance and its behaviour does not conflict with the exactly vanishing degree-one component in Corollary 3.3; it is reported because it is built from the same aligned marginals as the second-order term.

(α, p)	Statistic	n			
		15	25	40	60
$(0.6, 0.3)$	$\lvert S_1\rvert$ (folded, aligned)	83.75	81.35	83.63	79.48
	order 2, aligned	85.18	82.68	85.25	81.82
	order 2, generic	60.42	57.53	55.00	52.93
	order 3, aligned	74.89	71.65	73.71	70.93
$(0.8, 0.3)$	$\lvert S_1\rvert$ (folded, aligned)	72.27	72.28	69.56	67.19
	order 2, aligned	74.50	74.44	71.88	69.09
	order 2, generic	56.41	54.03	51.57	51.61
	order 3, aligned	65.59	66.76	62.44	61.26
$(0.6, 0.5)$	$\lvert S_1\rvert$ (folded, aligned)	85.28	84.06	85.22	80.34
	order 2, aligned	87.88	86.55	87.55	83.31
	order 2, generic	57.49	53.96	52.20	51.76
	order 3, aligned	77.60	74.90	78.48	73.30
$(0.8, 0.5)$	$\lvert S_1\rvert$ (folded, aligned)	73.08	74.70	72.32	69.76
	order 2, aligned	75.69	77.12	75.03	72.24
	order 2, generic	53.56	51.74	51.27	51.89
	order 3, aligned	67.11	69.10	66.93	65.33

E.7 Backbone sweep, resolved by size and density

Table E.5 groups cells post hoc by the reference GIN AUC, which answers whether the hard regime survives a change of aggregation rule but deliberately discards the (n, r) coordinates. The grouped view can therefore hide a backbone that is merely shifted along r rather than uniformly weaker. Table E.10 reports the same runs ungrouped, on the layout of the main GIN grid, so that each architecture can be read as its own transition. The sweep covers the $(\alpha, p) = (0.6, 0.3)$ grid with three seeds at each of the eight tabulated densities and all four sizes.

The ungrouped view reproduces the grouped conclusion and adds two details. All four architectures stay near chance at low density and improve as density increases, so the transition is a property of the instance family rather than of one aggregation rule, and none of them reaches the relation-level pair score at $n \in \{40, 60\}$, which is the regime the extraction gap concerns.

They diverge sharply in how far they get, and the divergence grows with size. The largest AUC attained anywhere in the sweep is 100.0, 100.0, 99.8 and 91.3 for GIN across $n = 15, 25, 40, 60$, and 100.0, 100.0, 99.7 and 93.6 for GraphSAGE. GAT matches them at the two smallest sizes and then falls away, to 84.4 at $n = 40$ and 58.5 at $n = 60$, while GCN is the weakest throughout, from 91.0 down to 56.2. At $n = 60$ the four models therefore split into two pairs rather than forming a gradient.

The split follows how each rule scales its neighbourhood aggregate. GCN normalises by degree and GAT by attention weights that sum to one, so both report a weighted average of the neighbourhood; GIN sums its neighbours and GraphSAGE keeps a separate root-weight path, so both retain an un-normalised magnitude. The evidence here is a count of redundant witnesses rather than a pattern in the composition of a neighbourhood, and the degree of a value node is exactly the number of constraints that forbid that value, so a normalising aggregator divides out the quantity that carries the signal. We record this as an observation consistent with the mechanism rather than a controlled result: it rests on four architectures on one grid, and aggregation is not the only thing that differs between them.

Table E.10: Backbone sweep on $(\alpha, p) = (0.6, 0.3)$, resolved by instance size and density, in percent (mean $\pm$ standard deviation over three seeds). Ungrouped companion to Table E.5.

Architecture	n	Metric	r/r_{cr} 0.5	0.75	1	1.5	2	3	4	6
GIN	15	Acc	52.83 ±1.04	55.17 ±0.62	57.92 ±1.45	71.71 ±0.50	81.67 ±0.75	95.83 ±0.31	98.79 ±0.47	99.88 ±0.00
		AUC	53.90 ±0.16	56.23 ±1.06	60.97 ±0.60	79.46 ±0.25	89.15 ±0.33	99.35 ±0.13	99.96 ±0.02	100.00 ±0.00
	25	Acc	50.63 ±1.39	52.63 ±2.13	53.66 ±2.15	62.71 ±1.96	69.08 ±1.56	82.04 ±1.41	92.29 ±0.29	99.04 ±0.19
		AUC	50.79 ±1.58	54.50 ±2.57	54.22 ±0.10	68.33 ±2.26	77.00 ±0.91	89.97 ±0.93	97.87 ±0.11	99.96 ±0.02
	40	Acc	50.44 ±0.92	52.55 ±0.63	50.33 ±2.78	53.44 ±1.27	60.95 ±1.55	74.00 ±0.72	85.78 ±1.57	97.67 ±0.17
		AUC	52.63 ±1.35	54.06 ±0.90	51.75 ±3.28	55.05 ±0.49	66.85 ±0.89	81.66 ±1.15	93.84 ±0.56	99.79 ±0.05
	60	Acc	49.67 ±0.76	49.08 ±2.63	49.58 ±1.04	51.92 ±0.88	50.83 ±0.38	54.75 ±1.75	68.92 ±2.04	83.58 ±1.66
		AUC	49.63 ±1.07	50.46 ±1.61	51.19 ±2.06	53.46 ±0.89	50.39 ±1.52	58.05 ±1.45	73.98 ±2.08	91.26 ±0.80
GAT	15	Acc	51.46 ±0.94	50.91 ±1.32	55.42 ±0.92	63.50 ±1.38	70.54 ±4.42	93.21 ±3.12	97.13 ±1.87	99.83 ±0.14
		AUC	52.64 ±0.32	51.23 ±0.62	56.35 ±1.56	69.68 ±3.27	78.55 ±4.19	98.50 ±1.12	99.64 ±0.34	100.00 ±0.00
	25	Acc	50.75 ±0.66	47.96 ±0.39	51.66 ±0.79	53.17 ±1.49	60.58 ±5.21	75.21 ±10.10	84.87 ±17.64	98.04 ±2.53
		AUC	52.18 ±0.49	48.76 ±0.42	51.78 ±0.20	53.75 ±1.53	64.13 ±7.20	82.56 ±12.29	88.43 ±18.75	99.66 ±0.41
	40	Acc	51.89 ±0.75	50.72 ±1.80	49.61 ±1.40	49.17 ±0.87	51.06 ±1.51	55.95 ±1.07	55.78 ±5.01	77.83 ±16.03
		AUC	52.89 ±0.28	51.05 ±0.46	50.11 ±2.87	48.34 ±0.96	50.80 ±0.52	56.23 ±2.18	58.86 ±5.87	84.39 ±18.28
	60	Acc	50.67 ±1.46	47.50 ±1.56	51.08 ±0.29	48.83 ±5.13	52.08 ±0.38	51.00 ±1.00	51.50 ±0.50	53.92 ±2.50
		AUC	50.16 ±0.65	45.40 ±0.66	51.69 ±0.26	48.05 ±4.72	51.30 ±0.23	53.22 ±1.07	52.60 ±0.53	58.50 ±1.11
GCN	15	Acc	52.25 ±1.07	51.83 ±0.62	52.54 ±2.93	55.83 ±0.80	59.75 ±2.22	66.79 ±0.47	72.54 ±1.02	83.00 ±0.12
		AUC	53.21 ±0.83	53.67 ±0.32	55.66 ±3.19	56.70 ±1.07	63.03 ±0.81	72.13 ±0.34	79.76 ±1.36	91.00 ±0.11
	25	Acc	50.58 ±0.69	50.63 ±0.55	51.50 ±0.25	50.79 ±0.97	56.00 ±0.21	61.88 ±0.38	59.71 ±0.71	70.29 ±0.26
		AUC	52.72 ±1.08	52.19 ±2.15	51.95 ±0.08	53.31 ±0.19	59.40 ±0.57	64.69 ±0.29	65.49 ±0.19	77.72 ±0.22
	40	Acc	50.11 ±0.19	49.61 ±0.54	51.39 ±1.42	50.17 ±0.17	50.00 ±0.00	54.89 ±1.73	53.22 ±2.80	64.00 ±0.29
		AUC	50.04 ±2.86	50.59 ±0.32	51.76 ±2.45	49.45 ±0.05	50.93 ±2.05	59.33 ±0.94	56.91 ±2.82	70.98 ±0.58
	60	Acc	50.33 ±1.04	50.25 ±0.43	50.25 ±0.43	51.92 ±2.50	50.08 ±0.14	50.00 ±0.00	50.00 ±0.00	54.83 ±5.97
		AUC	49.40 ±0.54	49.57 ±3.46	48.53 ±0.40	52.15 ±0.26	50.45 ±5.48	45.96 ±0.41	42.45 ±0.88	56.21 ±10.56
GraphSAGE	15	Acc	54.79 ±0.59	58.59 ±0.71	61.04 ±2.13	75.71 ±0.95	85.71 ±1.07	97.59 ±0.36	99.29 ±0.08	99.71 ±0.08
		AUC	57.58 ±0.18	64.19 ±0.23	66.80 ±2.42	83.22 ±1.28	93.67 ±0.32	99.75 ±0.08	99.99 ±0.01	100.00 ±0.00
	25	Acc	51.58 ±0.56	54.25 ±0.98	55.50 ±0.97	66.29 ±0.73	69.62 ±1.23	85.96 ±0.75	95.42 ±0.69	99.50 ±0.21
		AUC	54.83 ±1.25	58.15 ±1.14	58.67 ±0.92	71.90 ±1.12	77.61 ±0.57	92.31 ±0.43	99.32 ±0.14	100.00 ±0.01
	40	Acc	51.17 ±1.01	49.28 ±1.65	53.28 ±0.95	58.00 ±0.60	63.89 ±2.41	76.95 ±0.86	87.78 ±1.17	96.95 ±0.63
		AUC	52.47 ±1.43	48.23 ±2.06	54.82 ±0.81	61.72 ±0.06	69.57 ±1.45	85.92 ±0.35	94.85 ±1.04	99.73 ±0.08
	60	Acc	49.83 ±0.72	50.50 ±1.89	49.17 ±1.26	51.33 ±0.95	53.25 ±1.64	56.75 ±2.65	71.58 ±2.02	84.83 ±2.98
		AUC	50.39 ±1.27	49.38 ±0.68	52.80 ±2.03	50.67 ±1.25	53.22 ±1.00	60.93 ±0.38	78.16 ±1.60	93.59 ±1.58

E.8 Crossings beyond the swept density range

The empirical crossing $r_{\mathrm{GIN}}^{(0.70)}$ is obtained by interpolation and therefore exists only where the sweep reaches $\mathrm{AUC} = 0.70$. It does so in every $\alpha = 0.6$ cell and, for $\alpha = 0.8$, at $n \in \{15, 25\}$. At $\alpha = 0.8$ and $n \in \{40, 60\}$ the largest swept density still gives 54.1 and 49.1 for $p = 0.5$ and 54.8 and 54.2 for $p = 0.3$, so those four cells are reported as lower bounds and the boundary diagram marks them as exceeding the sweep.

On the two grids where all four sizes are available, the crossing follows a power law in problem size,

$$\frac{r_{\mathrm{GIN}}(n)}{r_{\mathrm{cr}}} \;\propto\; n^{\beta}, \tag{E.1}$$

with $\beta = 0.91$ at $(\alpha, p) = (0.6, 0.5)$ and $\beta = 0.77$ at $(0.6, 0.3)$. The fitted values $\{1.33, 2.13, 3.27, 4.73\}$ and $\{1.15, 1.70, 2.44, 3.33\}$ track the measured $\{1.33, 2.19, 3.10, 4.87\}$ and $\{1.24, 1.59, 2.14, 3.75\}$ to about 0.1. Fitting the same form to the two available sizes of each $\alpha = 0.8$ grid places the missing crossings near $9.0\, r_{\mathrm{cr}}$ at $n = 40$ for $p = 0.5$, an extrapolation from two points.

We tested that prediction directly by extending $(\alpha, p) = (0.8, 0.5)$ at $n = 40$ to $r/r_{\mathrm{cr}} \in \{7, 8, 9\}$, holding the training-set size, batch size, optimizer, schedule, seeds and evaluation protocol identical to the original $n = 40$ points. The curve stays between 48 and 54 AUC through $8\, r_{\mathrm{cr}}$ and then rises to 68.0 at $9\, r_{\mathrm{cr}}$, placing the measured AUC-0.65 crossing at 8.79 and the AUC-0.70 crossing just beyond 9, where Eq. (E.1) put it.

Three properties of these crossings bear on the theory. At fixed n the crossing is far larger for $\alpha = 0.8$ than for $\alpha = 0.6$, 3.43 against 1.33 at $n = 15$, which is the direction the second-order decomposition predicts, since enlarging the domain divides the evidence carried by each overlapping pair by $q^{-(2k-1)}$. The crossing also grows with n in every grid, as the theoretical scale does. The rates differ, however: Corollary 3.7 gives $r_{\mathrm{LD}}/r_{\mathrm{cr}} = \Theta(n^{(3\alpha-1)/2}/\log n)$, which over $n = 15 \to 60$ at $\alpha = 0.6$ grows by roughly 15%, whereas the measured $r_{\mathrm{GIN}}/r_{\mathrm{cr}}$ grows by a factor of about 3.7.

Table F.1: Finite configuration of the introductory case study. The exact finite-sample calculations use the realized integers and $p_{\text{eff}} = t/q^2$.

Quantity	Definition	Value
n	Number of variables	15
α	Domain exponent	0.6
p	Nominal tightness	0.5
r/r_{cr}	Density in threshold units	2.0
r_{cr}	$-\alpha/\log(1-p)$	0.865617
r	$2r_{\text{cr}}$	1.731234
q	$\text{round}(n^\alpha)$	5
m	$\text{round}(rn\log n)$	70
t	$\text{round}(pq^2)$, forbidden tuples per constraint	13
p_{eff}	t/q^2, realized tightness	0.52
$\lvert\mathcal{V}\rvert$	nq, microstructure-graph nodes	75

The learned boundary drifts much faster than the density at which the evidence becomes available, and that difference is the extraction gap of Section 4.3 seen in the density coordinate.

Densities below the threshold. Below r_{cr}, a null instance is satisfiable with high probability, so satisfiability itself does not asymptotically separate the classes. The two distributions nevertheless need not coincide: Λ_2 is positive for $m \geq 2$ and grows as shared-scope evidence accumulates. In our finite sweeps, detector and neural performance show an overall upward trend with r, with local fluctuations rather than exact monotonicity.

This detection problem differs from the classical search experiments on Model RB, where the cost of *finding* a satisfying assignment is concentrated near a satisfiability transition (Xu et al., 2007). Search cost, global satisfiability, low-degree visibility, and trained classification performance are therefore four distinct quantities.

F Reproducibility Details for the Introductory Case Study

This section specifies the experimental configuration, sampling procedure, label verification, and statistics used in Figure 2. The case study follows the finite Model RB conventions and graph construction described in Appendices D.1 and D.2.

F.1 Configuration and sampling

The case study uses the cell $(\alpha, p) = (0.6, 0.5)$ at $n = 15$ and $r = 2r_{\text{cr}}$. Table F.1 reports both the nominal parameters and their realized finite values.

Because $pq^2 = 12.5$ is rounded to $t = 13$, the realized tightness is $p_{\text{eff}} = 13/25 = 0.52$. All exact finite-sample quantities, including Λ_2, are evaluated using p_{eff} and the realized integers. The nominal parameter $p = 0.5$ is retained only for the asymptotic Model RB parameterization.

Each constraint scope is an independent uniform pair $\{i, j\}$ with $i \neq j$; sampling is with replacement, so repeated scopes are permitted. Conditional on its scope, a null constraint forbids a uniformly sampled size-t subset of $[q]^2$. A planted instance first samples $\sigma^* \sim \text{Unif}([q]^n)$ and then samples every forbidden set uniformly among the size-t subsets that exclude the planted tuple (σ_i^*, σ_j^*). The hidden assignment consequently satisfies every planted constraint by construction.

The two instances in Figure 2(a) use seed 7 and share the same scope realization, isolating the difference between null and planted relation sampling in the visualization. This pairing is used only for panel (a). Panel (b) contains 200 independently generated instances per class, with seeds $5000 + 101k + c$ for $k = 0, \ldots, 199$, where $c = 1$ denotes the planted class and $c = 2$ denotes the null class. The class-specific offsets prevent the two samples from sharing a random stream.

F.2 Label verification and reported statistics

Satisfiability can be expressed in either graph encoding. A satisfying assignment corresponds to an independent set of the microstructure conflict graph containing exactly one value node per variable, or equivalently to an n-clique in the complementary compatibility graph. We verify each displayed label using two exact procedures. Complete backtracking either returns a satisfying assignment or exhausts the search tree. Independently, exact maximum-clique enumeration computes the clique number of the compatibility graph. Both procedures certify that the planted instance is satisfiable and that the null instance is unsatisfiable. Their compatibility-graph clique numbers are 15 and 10, respectively.

The red edges in Figure 2(a) mark constraints violated by a full assignment returned by min-conflicts search. The displayed planted assignment violates no constraint, while the displayed null assignment violates 10 of 70 constraints. These red edges visualize one search outcome rather than a unique certificate. In particular, 10 is an upper bound on the minimum number of constraints violated by a full assignment and is not asserted to be optimal. The UNSAT label follows from the exact procedures above, not from the min-conflicts outcome. Similarly, the clique number 10 measures the largest consistent partial assignment and should not be identified with the number of violated constraints.

Figure 2(b) compares two instance-level statistics. The conflict count is the number of edges in the deduplicated microstructure graph, and thus counts distinct forbidden value pairs after graph construction. The shared-variable pair score is the statistic in Eq. (C.36), computed from centered relation marginals. At this single experimental cell, their AUC values are 0.523 and 0.936, respectively. These values provide an illustrative comparison rather than an estimate of the size-dependent transition. Sections 4.1 and 4.2 evaluate that transition over the full density grid, while Table E.2 reports the corresponding stratified results.

G Extended Related Work

Random CSPs, planted distributions, and Model RB. Random constraint-satisfaction problems provide a classical setting in which a single control parameter changes both typical instance structure and algorithmic behavior. Early empirical studies of random SAT connected the hardest search regimes to satisfiability phase transitions (Cheeseman et al., 1991; Mitchell et al., 1992), while later work established sharp threshold phenomena and identified barriers that accompany the associated change in solution geometry (Friedgut, 1999; Achlioptas & Coja-Oghlan, 2008). Planted constructions replace an unconditional random instance with one conditioned to contain a hidden witness. Depending on the model, marginalizing that witness can preserve low-order statistics, producing the quiet-planting phenomenon and making detection different from merely verifying a supplied solution (Krzakala & Zdeborová, 2009; Feldman et al., 2015).

Model RB was introduced to obtain random CSPs with polynomially growing domain size and an exact satisfiability threshold (Xu & Li, 2000). Its forced-satisfiable generator samples an assignment and conditions every relation to admit the corresponding local tuple (Xu et al., 2007). After the assignment is marginalized, this generator is precisely the planted distribution considered here. Prior analyses use Model RB primarily to characterize satisfiability and to generate satisfiable instances with controlled search difficulty. We instead use it as a controlled family of two-distribution graph-learning problems. The satisfiability threshold provides a global reference, while the planted assignment induces consistency across constraints without changing any one-constraint marginal.

This distinction connects Model RB to low-degree hypothesis testing, which projects a likelihood ratio onto bounded-degree polynomial spaces to quantify the evidence accessible to restricted statistics (Kunisky et al., 2019). Low-degree behavior can also predict the limits of statistical-query algorithms when the required regularity assumptions hold (Brennan et al., 2021). In this work, we derive the exact finite-degree likelihood-ratio norm for planted Model RB, identify overlapping constraint pairs as the first nonzero component, and prove logarithmic-degree indistinguishability below the resulting visibility scale. We use the low-degree conjecture only when discussing a possible extension to general polynomial-time computation. Our formal claims themselves concern polynomial statistics of the complete CSP observation.

Message passing, Weisfeiler–Leman, and higher-order graph learning. Graph neural networks originated as recursive models for graph-structured data and were later unified through the message-passing framework (Scarselli et al., 2009; Gilmer et al., 2017; Li et al., 2016; Battaglia et al., 2018). Widely used architectures instantiate this framework with different neighborhood operators, including normalized convolution, mean aggregation, attention, injective sum aggregation, and collections of multiple aggregators (Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018; Xu et al., 2019; Corso et al., 2020). Simplified propagation, multiscale skip connections, hierarchical pooling, permutation-invariant set aggregation, and learnable multiset pooling further change how local representations are combined into a graph-level prediction (Wu et al., 2019; Xu et al., 2018; Ying et al., 2018; Zaheer et al., 2017; Baek et al., 2021).

The Weisfeiler–Leman hierarchy gives a complementary account of structural expressivity. The one-dimensional test underlies classical graph kernels (Shervashidze et al., 2011) and characterizes the distinguishing power of broad classes of message-passing GNNs under standard assumptions (Xu et al., 2019; Morris et al., 2019). It cannot distinguish every pair of nonisomorphic graphs (Cai et al., 1992). Surveys and refined analyses clarify both this limitation and the distinctions among WL dimension, neural architecture, and task-dependent expressivity (Sato, 2020; Huang & Villar, 2021; Böker et al., 2023; Loukas, 2020b). Higher-order invariant and equivariant networks, relational pooling, and sparse tuple-based constructions increase the order of the maintained representation (Maron et al., 2019b; Keriven & Peyré, 2019; Azizian & Lelarge, 2021; Murphy et al., 2019; Morris et al., 2020). PPGN provides a practical pair-indexed instance of this family (Maron et al., 2019a), while homomorphism counts give a related mathematical characterization of structural distinguishability (Dvořák, 2010).

Another line of work enriches local computation with explicitly structural units. Representative approaches count selected substructures, aggregate rooted subgraphs, nest local GNNs, lift computation to cellular complexes, or propagate structural identifiers (Chen et al., 2020; Bouritsas et al., 2023; Bevilacqua et al., 2022; Zhang & Li, 2021; Bodnar et al., 2021; Vignac et al., 2020; Barceló et al., 2021; Frasca et al., 2022). These mechanisms may expose relational patterns that ordinary node-wise aggregation folds together, but they change different aspects of the learner and therefore should not be treated as interchangeable notions of "higher order."

The WL hierarchy orders graph-distinguishing power, but this order alone need not predict task performance. Homomorphism expressivity provides a more quantitative description of which patterns a model can count (Zhang et al., 2024). Complementarily, recent generalization analysis uses task-dependent graph pseudometrics to show that additional expressivity is useful only when it improves alignment with label-relevant structure at an appropriate statistical and computational cost (Maskey et al., 2026). That work studies labels constructed from an already specified graph statistic. Our setting asks an upstream question: when a planted global label is exactly absent from every single constraint marginal, at what density does redundant cross-constraint consistency first make that label visible, and which graph learning interfaces preserve and extract it?

Depth and global communication address a separate limitation. Deep message-passing networks can lose discriminative information through oversmoothing (Oono & Suzuki, 2020; Zhao & Akoglu, 2020; Rong et al., 2020), while restricted depth, width, and graph bottlenecks can prevent efficient long-range transport (Loukas, 2020a; Alon & Yahav, 2021; Topping et al., 2022). Diffusion-based propagation, graph transformers, and long-range benchmarks accordingly study how information can be transported or globally mixed (Gasteiger et al., 2019; Kreuzer et al., 2021; Rampášek et al., 2022; Dwivedi et al., 2022). These methods improve access to information that is present in the supplied graph, but communication alone cannot recover relation identity or multiplicity removed before learning.

In this work, the controllable Model RB family lets us vary the amount of redundant local evidence while preserving the meaning of the global label. We use this control to separate four questions that are often conflated: whether the instance contains low-order statistical evidence, whether the instance-to-graph map preserves that evidence, whether a representation class can express it, and whether finite-data training extracts it. GCN, GraphSAGE, GAT, and GIN compare local aggregation rules. Depth and a virtual node vary communication range. PPGN changes representation order. A relation-aware pair statistic retains the constraint interface. The resulting comparisons do not equate polynomial degree, GNN depth, and WL order. Instead, they locate which source of information changes when a learner begins to distinguish the two instance distributions.

H Implications for Local-to-Global Graph Learning

Model RB isolates a general local-to-global mechanism: a label can be defined by a global witness while remaining absent from each local observation, and it becomes visible only after overlapping observations provide sufficient redundant agreement. This perspective shifts attention from whether a graph learner can aggregate local information to whether the supplied local information preserves the relations that express the global semantics.

From local indistinguishability to global semantics. Quiet planting makes every individual constraint label-independent, while shared variables couple different constraints through the same hidden assignment. Constraint density therefore controls the redundancy connecting local observations to the global label. The resulting signal is neither a single local pattern nor the global witness itself. It is the repeated consistency among local objects. This intermediate relational structure explains why global semantics can remain invisible to local summaries and then emerge gradually as the instance becomes denser.

Communication and relational computation. The microstructure graphs in our experiments are already highly connected, and two aggregation rounds reach between 93% and 99% of their nodes. The relevant distinction is therefore not only how far information travels, but which interactions the aggregation retains. The planted signal is expressed through a particular contraction of relation tables associated with shared variables. Increasing depth or adding a global communication channel enlarges the receptive field, whereas relation-aware aggregation changes the quantity being computed. This distinction separates global access from global understanding.

Expressivity and task alignment. The Weisfeiler–Leman hierarchy orders architectures by structural separating power, but greater separating power need not emphasize the relations relevant to a particular label. In our experiments, relation-aligned features recover substantially more signal than the evaluated increase in generic expressivity. This observation agrees with analyses in which expressivity improves learning when it aligns with label-relevant structure (Maskey et al., 2026; Zhang et al., 2024). For local-to-global tasks, architecture design should therefore consider both the class of structures a model can separate and the relational summaries its representation makes accessible.

Global structure before local aggregation. These findings suggest incorporating suitable global context before or during local message passing. Hierarchical structural summaries, spectral or positional encodings, and factorized relation-aware features offer different ways to expose global organization without constructing a dense pair state. Structural entropy, for example, provides an encoding-tree decomposition of graph organization (Li & Pan, 2016), has been used to analyze real networks (Li et al., 2018), and can guide hierarchical graph pooling (Wu et al., 2022). Model RB supplies a controlled benchmark for comparing such summaries at matched Λ_2. This creates a path toward scalable graph learners that retain the efficiency of local aggregation while making globally meaningful relations available from the outset.