SAME MARGINALS, DIFFERENT QUERY PROBABILITIES: AUDITING RELATIONAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Equal low-order relational marginals need not determine a Boolean-query probability. We cast relational query prediction as an *information-sufficiency audit*: declare the world- and query-side information an estimator may use, construct task pairs that the declared interface cannot separate, compute their exact risk floors, and bracket the Bayes risk of each trained model's input between references that see less and more. We instantiate the audit with a benchmark whose even- and odd-parity mechanisms match every marginal through order three and whose query templates collide at three levels of query information. Coarse clause shape removes the support collision and lowers probability MSE by 0.033. On a crossed-pairing collision, the Bayes risk of every shape-only input lies in $[0.0206, 0.0211]$; a set model with named incidence and sampled worlds reaches 0.0048, matching the Bayes risk of incidence on the query edges. Training budget separates the two: the shape model stops inside its bracket, the incidence model keeps improving. The matched-marginal parity shift carries a large exploitable signal: with the true groups the parity-pool Bayes risk is 1.16×10^{-4}, against 6.45×10^{-4} for independent edges. Trained on the parity families alone, a fourth-moment MLP reaches 2.35×10^{-4} while the set model stays at the independent-edge value, and adding common shock to training erases the gain for both. An all-edge detector with AUROC 0.987 ranks query losses worse than ensemble disagreement.

1 INTRODUCTION

A relational predictor that observes finite possible worlds and returns a Boolean-query probability estimates a functional of the joint edge law, not of individual marginals (Dalvi & Suciu, 2004; Sen & Deshpande, 2007). Two laws can agree on every low-order marginal and still assign different probabilities to the same query. Whether a prediction is possible therefore depends on two things jointly: what the context reveals about the law, and what the query representation reveals about the requested functional.

We make this dependence operational with an *information-sufficiency audit*. The audit declares which world, query, and latent-structure information an estimator may use, builds task pairs that the declared interface maps to the same observation law but that have different targets, and computes the resulting exact risk floor (Le Cam, 1986; Manski, 2003). Nested interfaces then localize where a collision disappears, and trained models are measured against the floors of their own interface. The audit separates three sources of error that a single accuracy number conflates: information the interface discards, dependence the context cannot reveal, and learning that falls short of what the input supports.

Prior-data fitted networks amortize Bayesian inference over synthetic task priors (Müller et al., 2022; Hollmann et al., 2023) and now extend to graphs and relational databases (Eremeev et al., 2026; Wang et al., 2026; Kothapalli et al., 2026). These models are evaluated by predictive accuracy on held-out tasks. The audit asks a prior question: for a declared input contract, which task pairs can the input separate at all, and how close does a trained model come to the attainable risk?

Our benchmark crosses mechanism, query template, visible topology, interaction strength, missingness, and context size. Even and odd parity agree on all marginals through order three. Seven query

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.

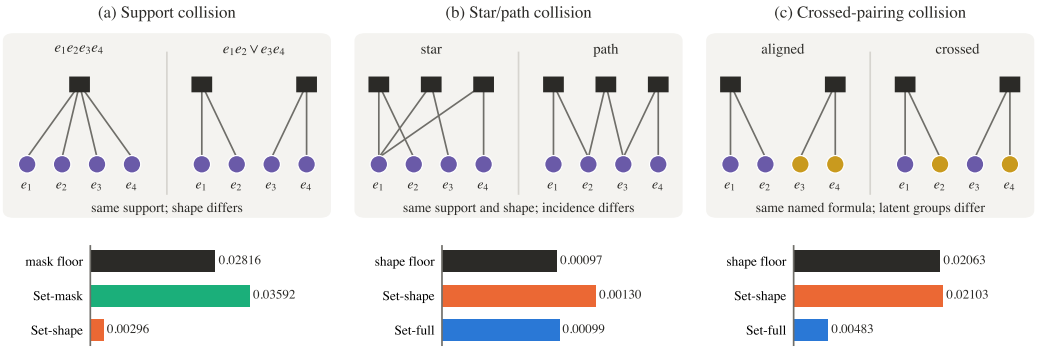


Figure 1: The three query collisions. Squares are clauses, circles are query edges, and circle colour marks the latent group. Each pair induces the same world law, so a predictor on the coarser interface pays the exact floor (black bar) even with mechanism and strength revealed. In (c) both templates are the named formula $e_1e_2 \vee e_3e_4$; only which edges co-vary in sampled worlds separates them. Bars give ensemble probability MSE from Table 4, averaged over mechanisms except in (c), which is common shock.

templates form three collisions with increasing query information: a conjunction and a union with the same support, a star and a path with the same clause shape, and two crossed pairings that are identical named formulas up to relabeling and are separated only by which edges share a latent group. Latent groups are drawn independently of visible topology. Our contributions are:

- **An identification audit with exact floors and Bayes brackets.** We define interface insufficiency, give two-point squared-error and interval-width bounds, compute exact floors for each collision, and bracket the Bayes risk of each learner's input between an oracle that sees more and a reference that reads a function of the same input (Section 3).
- **Trained evidence at every level of the query ladder.** Coarse shape removes the support collision. Named incidence beats shape on star/path tasks under every mechanism and falls below the shape floor on the training families. On crossed pairings, incidence with sampled worlds reaches 0.0048, the Bayes risk of its query-edge interface, against $[0.0206, 0.0211]$ for any shape-only input (Section 5.1).
- **Interface loss separated from learner excess.** Across a sixteen-fold range of training budgets, shape models stay inside their Bayes bracket while incidence models keep improving and overtake partition BMA (Section 5.2).
- **The matched-marginal signal is learnable, and the set learner misses it.** A full-information oracle shows that 82% of the independent-edge error on the parity pool is exploitable. Retraining under three priors splits the set learner's excess into prior mismatch and estimation error: a fourth-moment MLP trained on parity alone captures most of the margin, the set model captures none, and a mixed prior removes it for both (Section 5.3). A near-perfect mechanism detector ranks query losses worse than ensemble disagreement (Section 5.4).

2 Related Work

Probabilistic relational data and identification. Possible-world query evaluation and correlated tuples establish that query probabilities depend on the joint law beyond individual marginals (Dalvi & Suciu, 2004; Sen & Deshpande, 2007). We study the finite-context case in which mechanisms match low-order marginals yet imply different targets, connecting to statistical indistinguishability and partial identification (Le Cam, 1986; Manski, 2003). Parity and limited independence are standard Boolean tools (O'Donnell, 2014); we apply them to declared relational query interfaces. A monotone DNF is a semicoherent system whose minimal path sets are its clauses (Marichal, 2016), and reliability functions and system signatures admit established conversions (Marichal, 2015); our query interfaces coarsen this clause structure to its support or its clause-size profile.

Prior-fitted and relational inference. PFNs amortize Bayesian prediction over synthetic tasks (Müller et al., 2022; Hollmann et al., 2023). Related work studies temporal shift, runtime prior adaptation, and simulator misspecification (Helli et al., 2024; Whittle et al., 2026; Yang et al., 2026; Schmitt et al., 2024; Huang et al., 2023; Wehenkel et al., 2025) and extends synthetic-prior pretraining to graphs and relational databases (Eremeev et al., 2026; Wang et al., 2026; Kothapalli et al., 2026; Chen et al., 2026). Our set and incidence learners follow message-passing and permutation-invariant designs (Gilmer et al., 2017; Zaheer et al., 2017) at a scale where exact references are available for every task.

Logical-query representation. GQE, Query2Box, and BetaE embed logical-query operators (Hamilton et al., 2018; Ren et al., 2020; Ren & Leskovec, 2020), and provenance semirings preserve algebraic composition (Green et al., 2007). We quantify how much target information each coarsening of a query representation discards.

Distribution shift, selective prediction, and OOD detection. Label shift, two-sample testing, graph OOD benchmarks, and uncertainty under dataset shift study adjacent changes (Lipton et al., 2018; Gretton et al., 2012; Gui et al., 2022; Ovadia et al., 2019). Here low-order relational marginals stay fixed while higher-order dependence changes, so detecting the shift is distinct from predicting its query-specific consequence. Deep ensembles, selective classifiers, and AURC target predictive uncertainty or failure risk (Lakshminarayanan et al., 2017; Geifman & El-Yaniv, 2017; Corbière et al., 2019; Geifman & El-Yaniv, 2019; Zhou et al., 2025), and SCOD and failure-prediction studies separate OOD detection from selective risk (Narasimhan et al., 2024; Zhu et al., 2024). Conformal risk control offers related guarantees (Angelopoulos et al., 2024).

3 Problem Setup and Information-Sufficiency Audit

3.1 Tasks, targets, and declared interfaces

Let $\theta = (m, q)$ pair a possible-world mechanism with a monotone DNF query, with exact target $p(\theta) = p_q(m) = \Pr_m\{q(X) = 1\}$. The world-side observation is $O_m = (G, X_{\text{obs}}^{1:W})$, where G is a typed 16-edge graph and the context holds W partially observed worlds. An estimator's declared interface is

$$Z_{\psi,\phi}(\theta) = \big(\psi(O_m), \phi(q)\big), \tag{1}$$

where ψ controls world-side context and ϕ controls query-side representation. The evaluated interfaces form two enrichment chains, $\psi_\emptyset \prec \psi_{\text{worlds}}$ and $\phi_{\text{support}} \prec \phi_{\text{shape}} \prec \phi_{\text{incidence}}$. Here ϕ_{support} keeps the set of named edges, ϕ_{shape} adds the clause count and sorted clause sizes, and $\phi_{\text{incidence}}$ keeps named clause–edge membership. An interface is *insufficient* for a task pair (θ_0, θ_1) if

$$\mathcal{L}_{\theta_0}(Z_{\psi,\phi}) = \mathcal{L}_{\theta_1}(Z_{\psi,\phi}) \quad \text{and} \quad \Delta(\theta_0, \theta_1) := |p(\theta_0) - p(\theta_1)| > 0. \tag{2}$$

Insufficiency turns a target gap into a risk floor that no estimator on the interface can beat, however it is trained.

Proposition 1 (Two-point interface bound)**.** *Let θ_0, θ_1 induce the same law of the available information O and have targets p_0, p_1 with $\Delta = |p_1 - p_0|$. For any estimator $T = T(O, U)$ whose randomization U follows a common kernel $K(du \mid O)$ under both tasks, $\max_i \mathbb{E}_{\theta_i}(T - p_i)^2 \geq \Delta^2/4$. Any interval $I(O, U)$ with $\Pr_{\theta_i}\{p_i \in I\} \geq 1 - \alpha$ for $i \in \{0, 1\}$ satisfies $\mathbb{E}|I| \geq (1 - 2\alpha)_+\Delta$.*

For two equally weighted tasks, the minimum average squared error of a common prediction is also $\Delta^2/4$, attained at the midpoint; averaging it over a task grid gives the collision floors below. Proofs are in Appendix A.

We score prediction $\hat{p}$ by task-level probability MSE, $(\hat{p}_q - p_q(m))^2$, which equals expected Brier loss minus the irreducible outcome variance $p_q(1 - p_q)$.

Interface loss and learner excess. For a fixed task distribution and a predictor f that depends on the task only through its input Z, $\mathbb{E}[(f(Z) - p)^2] = R^*(Z) + \mathbb{E}[(f(Z) - \mathbb{E}[p \mid Z])^2]$ with $R^*(Z) = \mathbb{E}[\text{Var}(p \mid Z)]$. For a rich reference $Z_r = (O_m, \phi_{\text{incidence}}(q))$ and a coarse interface Z_c that is a function of Z_r,

$$R(f_c) = \underbrace{R^*(Z_r)}_{\text{rich residual}} + \underbrace{R^*(Z_c) - R^*(Z_r)}_{\text{interface loss}} + \underbrace{R(f_c) - R^*(Z_c)}_{\text{learner excess}}. \tag{3}$$

Collision floors computed with the mechanism and strength revealed lower-bound $R^*(Z_c)$, and any predictor that reads only Z_c upper-bounds it; we report both where available. When the training prior differs from the evaluation population, learner excess splits further into prior mismatch, $R(f_c) - R(f_c^\pi)$, and estimation error, $R(f_c^\pi) - R^*(Z_c)$, where f_c^π is the same learner trained on the evaluation prior. A model that settles between the two bounds of its interface is information-limited; one that improves with training or with a matched prior is learning-limited (Section 5.2).

3.2 World-side mechanisms and the matched-moment shift

Edges are partitioned into latent groups of four. For parity sign $s \in \{0, 1\}$ and $0 \le \lambda < 1$,

$$P_{s,\lambda}(x) = \frac{1-\lambda}{16} + \frac{\lambda}{8}\mathbf{1}\{\mathrm{parity}(x) = s\}. \tag{4}$$

Proposition 2 (Matched marginals)**.** *Under Equation 4, every marginal on a strict subset of a group's four edges is uniform for both parity signs, and for one fully observed group*

$$D_{\mathrm{KL}}(P_{\mathrm{even},\lambda} \,\|\, P_{\mathrm{odd},\lambda}) = \lambda \log \frac{1+\lambda}{1-\lambda}. \tag{5}$$

Even and odd parity therefore agree on all one-, two-, and three-edge marginal laws and differ only in the Rademacher fourth moment $\mu_4 = \mathbb{E}\prod_{i=0}^{3}(1-2X_i) = \pm\lambda$. Under independent missingness at rate r only complete group observations carry this signal, scaling it by $(1-r)^4$. A common-shock family uses the same uniform component and places its special mass on 0000 and 1111; it changes pairwise dependence. Training uses common shock and even parity; odd parity supplies the deployment shift. Section 5.3 retrains under a three-family and a parity-only prior to separate prior mismatch from estimation error.

3.3 Query templates and three collisions

Table 1 lists the seven templates with their exact targets. Groups A and B are distinct latent groups; the generator places query edges relative to these hidden groups and then applies random edge labels.

A mechanism change need not change a target: for even versus odd parity, the sensitivity $\Delta_q = |p_q(m_0) - p_q(m_1)|$ is $\lambda/8$ for the four-edge conjunction, the two-pair union, and the star, and zero for the other four templates. Three template pairs are indistinguishable at successive query interfaces, because each pair has the same support edges and the same latent-group structure and so induces the same world law:

- **Support collision.** The four-edge conjunction and the two-pair union share support but not shape. Averaging $\Delta^2/4$ over the mechanism–strength grid gives a support-only floor of $721/25600 = 0.02816$; shape separates the pair.
- **Star/path collision.** The star and the path share support and shape but not incidence. The shape-only floor is $33/102400$ under common shock or even parity and $233/102400$ under odd parity, averaging $299/307200 \approx 0.00097$. Named incidence separates the pair without any worlds; the remaining uncertainty is about mechanism and strength.
- **Crossed-pairing collision.** The aligned and crossed unions share support and shape; the gap is $\lambda/2$ under common shock and zero under parity, giving a shape-only floor of $33/1600 = 0.0206$ on common-shock tasks. Because edge labels are random, the two named formulas are identical up to relabeling: named incidence separates the pair only when sampled worlds reveal which named edges co-vary.

Separating a pair removes its two-point floor; the target is still not point-identified, since mechanism and strength remain uncertain under any finite context.

3.4 Factorial design

Visible topology is a path, hub, cycle, or set of disjoint edges, and latent groups are assigned by an independent random permutation, so adjacency carries no information about dependency groups. Mechanism, template, topology, strength $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, missingness $r \in \{0, 0.2, 0.4\}$,

Table 1: Query templates and exact targets. $a = (1 + \lambda)/4$ is the within-group pair probability under common shock. Shape lists clause sizes; all templates except the three-edge conjunction have support size four.

Template	Shape	Common shock	Even parity	Odd parity
$A_0A_1A_2$	$\{3\}$	$(1+3\lambda)/8$	$1/8$	$1/8$
$A_0A_1A_2A_3$	$\{4\}$	$(1+7\lambda)/16$	$(1+\lambda)/16$	$(1-\lambda)/16$
$A_0A_1 \vee A_2A_3$	$\{2,2\}$	$(7+\lambda)/16$	$(7-\lambda)/16$	$(7+\lambda)/16$
$A_0A_1 \vee A_0A_2 \vee A_0A_3$ (star)	$\{2,2,2\}$	$(7+\lambda)/16$	$(7+\lambda)/16$	$(7-\lambda)/16$
$A_0A_1 \vee A_1A_2 \vee A_2A_3$ (path)	$\{2,2,2\}$	$1/2$	$1/2$	$1/2$
$A_0A_1 \vee B_0B_1$ (aligned)	$\{2,2\}$	$2a - a^2$	$7/16$	$7/16$
$A_0B_0 \vee A_1B_1$ (crossed)	$\{2,2\}$	$1/2 - a^2$	$7/16$	$7/16$

and context size $W \in \{8, 24, 64\}$ are fully crossed. Within each paired design seed, common-shock and even-parity tasks share the query, topology, edge order, relabeling, and missingness mask; odd-parity tasks use a separate seed namespace and are matched by factorial cell. Training, calibration, and test partitions use disjoint seed namespaces (Appendix B).

4 Estimators, Cohorts, and Inference

4.1 Declared information boundaries

Table 2 records each estimator's input. Neural models read all 16 edges in every world; the non-oracle statistical references read only query edges. Only the two oracles see latent groups.

Neural interfaces. Set-mask, Set-shape, and Set-full share one 27,081-parameter backbone and differ only in the query encoder. Each world produces edge-state embeddings, the query encoder pools them within and across clauses in that world, and a one-layer Transformer aggregates the world tokens, so the full model can pair a clause with the simultaneous states of its edges. Shape spreads each sorted clause size uniformly over the support; mask uses one support clause. Incidence models add message passing along visible adjacency; a width-64 Set-full has 105,361 parameters. The moment MLP receives query-structure summaries and query-support moments up to fourth order; the structure MLP drops every world-derived feature (Appendix C).

Statistical references. Independent edges evaluate the query under Bernoulli$(1/2)$ edges, which is the exact marginal of every mechanism. Partition BMA averages over every unlabeled partition of the query edges (5 for three edges, 15 for four), mechanism families, and a strength grid, with a likelihood that sums out missing coordinates. The supported version uses the training families; the all-family version adds odd parity; an off-grid variant excludes every true strength. Two privileged references receive the true latent groups. The query-group oracle reads all four edges of each group the query touches. The full-information oracle reads every edge of every group, whose shared mechanism and strength make non-query groups informative, and integrates over the exact benchmark prior; its risk is the Bayes risk of worlds plus incidence plus true groups.

4.2 Cohorts

Cohorts A–D train on common-shock and even-parity tasks and evaluate on held-out tasks from all three mechanisms (Table 3). Cohorts P and T reuse the data and model seeds and the test tasks of the first five runs of cohorts B and D but change the training families. One epoch is one pass over 32,768 training tasks at batch size 64. Budget curves save predictions at 1, 2, 4, 8, and 16 epochs of one constant-learning-rate trajectory per run.

4.3 Metrics and inference

Point prediction is probability MSE against exact targets; networks train with soft-target cross-entropy. We report ten-run (cohorts C, P, T: five-run) ensemble risk and paired differences between interfaces trained with shared seeds. Intervals are 2,000-draw crossed bootstraps that resample training runs and

Table 2: Estimator inputs. Shape gives support, clause count, and sorted clause sizes; incidence gives named clause–edge membership.

Estimator	Worlds	Query	Mechanism families	Latent groups
Set-mask / Set-shape / Set-full	all edges	support / shape / incidence	learned	no
Incidence-full	all edges + adjacency	incidence	learned	no
Moment MLP; structure MLP	query-edge moments; none	incidence summaries	learned	no
Independent edges	none	incidence	Bernoulli$(1/2)$	no
Partition BMA (supported / all families)	query edges	incidence	common, even / + odd	marginalized
Query-group oracle	all edges of query groups	incidence	common, even, odd	**true**
Full-information oracle	all edges	incidence	exact benchmark prior	**true**

Table 3: Evaluation cohorts. Templates are those used in training; cohorts D and T are also tested on star and path tasks. Moment and structure MLPs train on the base templates with the seeds of cohort B, under each of the three priors. Test tasks are disjoint from training tasks and from each other.

Cohort	Templates	Test tasks	Runs	Epochs	Used for
A: frozen	4 base	4,320	10	1	topology, detection, and risk ranking
B: base budget	4 base	4,320	10	1–16	support collision, budget, parity
C: same-shape	7	3,780	5	1–16	crossed pairing, all collisions
D: star/path	4 base + star, path	4,320	10	1–16	star/path collision
P: parity prior	4 base; even, odd only	4,320	5	8	estimation error
T: three-family	as D; all three families	4,320	5	8	prior mismatch

a shared task vector; a fixed-grid paired-block bootstrap that resamples replicates inside each factorial cell gives a sensitivity analysis. For reliability scores, AUROC measures odd-parity detection and tie-expected AURC measures how well a confidence score ranks the ensemble's query loss (Appendix C).

5 Results

5.1 Trained interfaces follow the query-information ladder

Table 4 compares each collision's exact floor with the trained ensembles on both sides of it. Each row changes only the query encoder of a shared Set backbone.

Support to shape. On the conjunction/union pair, the support mask stays above its 0.02816 floor, and shape removes almost all of that error. Adding incidence then makes the model slightly worse, $+0.00056$ at one epoch and $+0.00018$ at sixteen, although both interfaces have a zero floor here. This is learner excess in the sense of Equation 3: the richer encoder must learn to ignore clause–edge detail that carries no information on this pair, and the cost shrinks with training.

Shape to incidence. The star/path pair is the smallest collision in the benchmark. Averaged over mechanisms, Set-full reaches 0.00099 after sixteen epochs, at the 0.00097 floor, and Set-shape stays above it at 0.00130. Incidence beats shape under every mechanism and both training priors (Table 14); it goes below the shape floor only on the training families, where width-64 Set-full reaches 0.000278 against $33/102400 = 0.000322$ (difference -0.000044, $[-0.000056, -0.000022]$). Training on all three families (cohort T) brings odd parity to its floor (0.00225 against 0.00228) and moves the training families above theirs. The query itself carries most of this information: a world-free MLP on named structure reaches 0.000157 on the training families, below Set-full's 0.000305, so Set-full carries learner excess here. All-family partition BMA reaches only 0.000375 because its uniform prior over the 15 partitions of four edges does not match the generator, which keeps star and path edges in one group; with the true group it reaches 0.000113.

Incidence with context. Crossed pairings isolate the case that requires both sides of the interface. The shape model settles at 0.0210 from the second epoch onward. The Bayes risk of any shape-only input lies in $[0.0206, 0.0211]$: the floor with mechanism and strength revealed is the lower end, and the exact Bayes predictor that reads the query edges of every world together with the clause shape is the upper end. Set-shape sits inside this bracket, so its remaining error is interface loss. The incidence model reaches 0.0048, four times below any shape-only predictor (Figure 2a), and equal to the exact

Table 4: Collision floors and trained risk on the two templates of each pair. The floor binds every predictor with the coarse interface even with mechanism and strength revealed. Risks are probability MSE, given as ensemble / mean single run; differences are paired ensemble contrasts (rich minus coarse) with crossed run–task 95% intervals. Rows average common shock, even, and odd parity with equal weight, except the crossed row, which is common shock only because the pair has no gap under parity.

Collision	Coarse → rich	Cohort, epochs	Coarse floor	Coarse risk	Rich risk	Difference [95% CI]
Support	mask → shape	B, 1	0.02816	0.03592 / 0.03721	0.00296 / 0.00322	-0.03296 [-0.03620, -0.03038]
Support	shape → incidence	B, 1	0	0.00296 / 0.00322	0.00352 / 0.00374	$+0.00056$ [$+0.00033$, $+0.00076$]
Support	shape → incidence	C, 16	0	0.00185 / 0.00189	0.00203 / 0.00213	$+0.00018$ [$+0.00006$, $+0.00032$]
Star/path	shape → incidence	D, 16	0.00097	0.00130 / 0.00133	0.00099 / 0.00101	-0.00032 [-0.00037, -0.00028]
Crossed	shape → incidence	C, 16	0.02063	0.02103 / 0.02104	0.00483 / 0.00509	-0.01619 [-0.01821, -0.01395]

Bayes risk of named incidence with query-edge worlds (0.0047). The two unions are the same named formula up to relabeling, so the model must infer from co-occurrence in the worlds which named edges share a group. The full-information oracle reaches 0.0003 (Table 13); Set-full reads the edges outside the query but does not exploit them here.

5.2 Budget separates interface loss from learner excess

Equation 3 predicts different budget responses for information-limited and learning-limited interfaces, and Figure 2 shows both. On crossed pairings, Set-shape reaches its floor within two epochs and stops; its remaining error is interface loss. Set-full keeps improving, from 0.0177 at one epoch to 0.0048 at sixteen; its one-epoch error was learner excess.

On the four base templates, Set-full deployment risk falls from 0.00360 to 0.00175 between one and sixteen epochs and ends below Set-shape (-0.00034, $[-0.00046, -0.00023]$). The same training reverses the comparison with all-family partition BMA: Set-full trails it by $+0.00112$ ($[+0.00093, +0.00139]$) at one epoch and leads by -0.00074 ($[-0.00094, -0.00051]$) at sixteen. The gain is on common shock (Figure 2c), where Set-full reaches 0.00236 against 0.00630 for BMA, 0.00207 for the query-group oracle, and 0.00064 for the full-information oracle, and is still falling: all groups share the task's mechanism and strength, so edges outside the query inform the target, and Set-full reads them while BMA does not.

5.3 The matched-marginal signal is learnable; the set learner misses it

Parity changes only the fourth-order interaction, so independent edges are exact on every parity-insensitive template and miss the two sensitive base templates by $\lambda/16$; their risk is $\frac{1}{2}\mathbb{E}[\lambda^2]/256 = 0.000645$ under either sign. Table 5 measures how much of this error the sampled worlds can remove. The full-information oracle reaches 0.000116, so 82% of the independent-edge error is exploitable. The query-group oracle, which reads only the groups the query touches, reaches 0.000428 and recovers less than half of that margin: all four groups share the task's mechanism and strength, and the twelve edges outside the query carry most of the evidence about them.

Learner excess. The set learners read all sixteen edges but not the groups. Partition BMA is a function of their input and the full oracle sees strictly more, so the Bayes risk of the Set-full interface lies in $[0.000116, 0.000571]$. Set-full attains 0.001436 after sixteen epochs: its learner excess is at least 8.7×10^{-4}, and it stays above independent edges even on even parity, a training family ($+0.00027$, $[+0.00007, +0.00053]$). Retraining under two other priors splits this excess in the two parts of Equation 3.

Prior mismatch. Odd parity lies outside the common/even training families. Adding it (cohort T) lowers single-model odd-parity risk from 0.00207 to 0.00074 and raises even-parity risk from 0.00092 to 0.00118, since each family receives fewer training tasks. About 1.3×10^{-3} of the odd-parity error is prior mismatch. Cohort T also trains on star and path templates, so this comparison uses single-model point estimates.

Estimation error. Trained on even and odd parity alone, the moment MLP, which receives the empirical fourth moment of the query support, reaches 0.000235. That is 0.00041 below independent

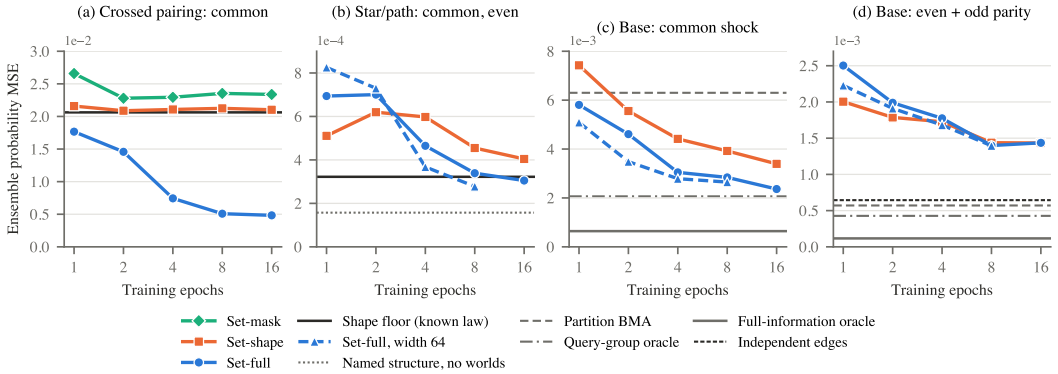


Figure 2: Risk against training budget. (a) Crossed pairing under common shock (cohort C): shape stays at its exact floor while incidence falls far below it. (b) Star/path on the training families (cohort D). (c–d) Four base templates (cohort B): incidence models overtake partition BMA under common shock, and every learner stays above independent edges on the parity pool, far from the full-information oracle. Horizontal lines are exact or deterministic references.

Table 5: Strict even/odd parity pool on the cohort-B test tasks, grouped by training prior. References in the last block use the parity-only prior. Differences are paired ensemble contrasts against independent edges with 95% intervals.

Prior	Predictor	Even	Odd	Pool	Minus independent [95% CI]
benchmark (Bayes refs.)	Independent edges (no worlds)	0.000645	0.000645	0.000645	–
	Partition BMA, all families	0.000524	0.000618	0.000571	−0.000074 [−0.000143, +0.000010]
	Query-group oracle	0.000409	0.000446	0.000428	–
	Full-information oracle	0.000112	0.000120	**0.000116**	–
common, even	Set-full, 16 epochs, 10 runs	0.000912	0.001960	0.001436	+0.00079 [+0.00064, +0.00098]
	Moment MLP	0.001271	0.002339	0.001805	+0.00116 [+0.00095, +0.00140]
common, even, odd	Set-full (cohort T), 8 epochs	0.001071	0.000654	0.000862	+0.00022 [+0.00013, +0.00037]
	Moment MLP	0.000977	0.001049	0.001013	+0.00037 [+0.00022, +0.00055]
even, odd	Partition BMA	0.000343	0.000349	0.000346	–
	Query-group oracle	0.000168	0.000187	0.000177	–
	Full-information oracle	0.000062	0.000067	0.000065	–
	Set-full (cohort P), 8 epochs	0.000745	0.000564	0.000654	+0.000010 [+0.000004, +0.000021]
	Moment MLP	0.000218	0.000252	**0.000235**	−0.00041 [−0.00044, −0.00037]
	Structure MLP (no worlds)	0.000632	0.000658	0.000645	–

edges, below partition BMA under the same prior (0.000346), and 70% of the way from independent edges to the parity-prior full oracle (0.000065). Set-full trained on the same tasks stays at the independent-edge value (0.000654). The moment features are a function of Set-full's input, so under the parity prior the Bayes risk of that input is at most 0.000235, and Set-full's estimation error is at least 4.2×10^{-4}. The signal is learnable from finite contexts; the set architecture does not extract it.

The mixed prior explains why no learner uses the signal under the benchmark prior. With common shock added to the training mixture, the moment MLP's parity-pool risk rises to 0.001013, above independent edges. Common-shock errors are an order of magnitude larger than the parity margin and dominate the squared-error objective, so a learner fitted to the mixture spends its capacity there. Whether the matched-marginal signal is captured depends on the composition of the training prior and on the architecture, not on the size of the signal.

5.4 Detection and risk ranking respond to different information

On the frozen deployment mixture (cohort A), partition GOF and the observed fourth moment detect odd parity with AUROC 0.725 and 0.783 (the latter reliably above about 3 nats of complete-group information; Figure 3), and ensemble variance with 0.440. For ranking query loss the order reverses: ensemble confidence has AURC 0.00332, against 0.00591 for GOF and 0.00488 for the fourth moment. Ensemble minus fourth moment is −0.00156 ([−0.00279, −0.00035]) on the

deployment mixture and $+0.00028$ ($[-0.00049, +0.00075]$) on the even/odd pool; the difference between the two populations, -0.00184 ($[-0.00285, -0.00092]$), is resolved under fixed-grid paired-block resampling.

Table 6: Ensemble-minus-fourth-moment AURC on cohort A by template (fixed-grid paired-block 95% intervals; negative favours the ensemble). Δ_q is the even/odd target gap.

Stratum	Δ_q	Deployment mixture	Even/odd pool
Four-edge queries only		-0.00188 $[-0.00332, -0.00039]$	$+0.00016$ $[-0.00045, +0.00076]$
Cross-group union	0	-0.00638 $[-0.00903, -0.00131]$	$+0.00015$ $[-0.00063, +0.00067]$
Three-edge conjunction	0	-0.00025 $[-0.00145, +0.00027]$	$+0.00043$ $[-0.00108, +0.00100]$
Two-pair union	$\lambda/8$	-0.00064 $[-0.00120, -0.00027]$	-0.00100 $[-0.00201, -0.00022]$
Four-edge conjunction	$\lambda/8$	$+0.00295$ $[-0.00114, +0.00374]$	$+0.00690$ $[-0.00026, +0.00864]$

Table 6 locates the difference. The fourth moment scores every three-edge query as zero, but that is not the source of the ensemble's advantage: on four-edge queries alone it is -0.00188 ($[-0.00332, -0.00039]$). The advantage is largest on the cross-group union, whose target does not depend on parity, and it shrinks or reverses on the parity-sensitive templates. Which score ranks query loss better follows how strongly each template's target responds to the mechanism change.

Giving the detector all the information sharpens the separation. The full-information oracle's posterior probability of odd parity detects odd parity with AUROC 0.987 on the deployment mixture and 0.982 on the even/odd pool. As a ranking score it is worse than ensemble confidence on the deployment mixture: AURC 0.00560 against 0.00332, a difference of -0.00228 ($[-0.00447, -0.00079]$). On the even/odd pool the two are not resolved ($+0.00061$, $[-0.00013, +0.00111]$). On the deployment mixture the largest query losses are on common-shock tasks (0.0066 against 0.0035 on odd parity; Table 11), and a parity detector admits exactly those tasks first. Detecting a mechanism change and ranking its consequence for a query are different tasks.

Visible topology, a negative control since groups are drawn independently of adjacency, adds nothing: shuffling the adjacency changes risk by -0.000028 ($[-0.000100, +0.000032]$; Table 16).

6 Conclusion

An information-sufficiency audit asks, before any accuracy is reported, which task pairs a declared interface can separate and what risk it can attain. On a benchmark with exact targets, each level of the query ladder removes the collision it should, and incidence with sampled worlds reaches the Bayes risk of its query-edge interface on a collision that no shape-only input can resolve. Budget curves separate models inside the Bayes bracket of their input from models that are still learning. The parity shift is invisible to every marginal through order three, yet 82% of the independent-edge error is exploitable; a fourth-moment learner trained on the parity families recovers most of it, and the set learner recovers none.

Three practices follow for evaluating relational probability estimators. First, report the input contract together with a Bayes bracket for it: a lower bound from a reference that sees more, and an upper bound from any predictor that reads a function of the input. The lower reference must see the whole context; an oracle restricted to the query's own groups understated the parity margin by a factor of 2.4. Second, report risk across training budgets and training priors. A single budget can reverse the ranking between a learned model and a statistical reference, and a single prior conflates prior mismatch with estimation error. Third, measure the value of a shift separately from its detectability: an all-edge detector with AUROC 0.987 ranks query losses worse than ensemble disagreement.

The benchmark uses sixteen edges and four latent groups so that every reference is exact. Auditing larger graphs, richer query languages, and pretrained relational foundation models against bounded rather than exact references is the next step.

Reproducibility Statement

Appendix B lists every cohort, seed namespace, and training budget; Appendix A proves the interface bounds. The anonymous supplement contains the generator, training and analysis code, tests, configurations, and prediction-level outputs for every reported number.

AI USE STATEMENT

Generative AI tools assisted language editing, code development, and statistical reanalysis. All numerical results are produced by the released code and tested against exact enumeration.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Zhikai Chen, Junyu Yin, Jialiang Gu, Siheng Xiong, Xiaoze Liu, Ruowang Zhang, Keren Zhou, and Kai Guo. OpenRFM: Dissecting relational in-context learning. *arXiv preprint arXiv:2606.04320*, 2026. URL https://arxiv.org/abs/2606.04320.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/757f843a169cc678064d9530d12a1881-Abstract.html.

Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pp. 864–875, 2004. URL https://www.vldb.org/conf/2004/RS22P1.PDF.

Dmitry Eremeev, Oleg Platonov, Gleb Bazhenov, Artem Babenko, and Liudmila Prokhorenkova. GraphPFN: A prior-data fitted graph foundation model. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.21489.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

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. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 31–40, 2007. doi: 10.1145/1265530.1265535. URL https://www.cs.ucdavis.edu/~green/papers/pods07.pdf.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0dc91de822b71c66a7f54fa121d8cbb9-Abstract-Datasets_and_Benchmarks.html.

William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. Embedding logical queries on knowledge graphs. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/ef50c335cca9f340bde656363ebd02fd-Abstract.html.

Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient TabPFN: In-context learning temporal distribution shifts on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b2e2774c8e76afe191b5bf518f5cb727-Abstract-Conference.html.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cp5PvcI6w8_.

Daolang Huang, Ayush Bharti, Amauri H. Souza, Luigi Acerbi, and Samuel Kaski. Learning robust statistics for simulation-based inference under model misspecification. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/16c5b4102a6b6eb061e502ce6736ad8a-Abstract-Conference.html.

Vignesh Kothapalli, Rishabh Ranjan, Valter Hudovernik, Vijay Prakash Dwivedi, Johannes Hoffart, Carlos Guestrin, and Jure Leskovec. PluRel: Synthetic data unlocks scaling laws for relational foundation models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.04029.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York, 1986. doi: 10.1007/978-1-4612-4946-7. URL https://link.springer.com/book/10.1007/978-1-4612-4946-7.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003. doi: 10.1007/b97478. URL https://link.springer.com/book/10.1007/b97478.

Jean-Luc Marichal. Algorithms and formulas for conversion between system signatures and reliability functions. *Journal of Applied Probability*, 52(2):490–507, 2015. doi: 10.1017/S0021900200012596. URL https://arxiv.org/abs/1312.7456.

Jean-Luc Marichal. Structure functions and minimal path sets. *IEEE Transactions on Reliability*, 65 (2):763–768, 2016. doi: 10.1109/TR.2015.2513017. URL https://arxiv.org/abs/1401.0803.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d78d68cae595fabadd187b583ee8708e-Abstract-Conference.html.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Hongyu Ren and Jure Leskovec. Beta embeddings for multi-hop logical reasoning in knowledge graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e43739bba7cdb577e9e3e4e42447f5a5-Abstract.html.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2Box: Reasoning over knowledge graphs in vector space using box embeddings. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJgr4kSFDS.

Marvin Schmitt, Paul-Christian Bürkner, Ullrich Köthe, and Stefan T. Radev. Detecting model misspecification in amortized bayesian inference with neural networks. In *Pattern Recognition: 45th DAGM German Conference, GCPR 2023, Proceedings*, volume 14264 of *Lecture Notes in Computer Science*, pp. 541–557. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-54605-1_35. URL https://doi.org/10.1007/978-3-031-54605-1_35.

Prithviraj Sen and Amol Deshpande. Representing and querying correlated tuples in probabilistic databases. In *Proceedings of the 23rd International Conference on Data Engineering*, pp. 596–605, 2007. doi: 10.1109/ICDE.2007.367905. URL https://doi.org/10.1109/ICDE.2007.367905.

Yanbo Wang, Jiaxuan You, Chuan Shi, and Muhan Zhang. Relational in-context learning via synthetic pre-training with structural prior. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2603.03805.

Antoine Wehenkel, Juan L. Gamella, Ozan Sener, Jens Behrmann, Guillermo Sapiro, Joern-Henrik Jacobsen, and Marco Cuturi. Addressing misspecification in simulation-based inference through data-driven calibration. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65949–65980, 2025. URL https://proceedings.mlr.press/v267/wehenkel25a.html.

George Whittle, Juliusz Ziomek, Jacob Rawling, and Maike A. Osborne. Distribution transformers: Fast approximate bayesian inference with on-the-fly prior adaptation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2502.02463.

Yang Yang, Severi Rissanen, Paul E. Chang, Nasrulloh Ratu Bagus Satrio Loka, Daolang Huang, Arno Solin, Markus Heinonen, and Luigi Acerbi. PriorGuide: Test-time prior adaptation for simulation-based inference. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=G4I23g5Ugh.

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. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B. Blaschko. A novel characterization of the population area under the risk coverage curve (AURC) and rates of finite sample estimators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79226–79253, 2025. URL https://proceedings.mlr.press/v267/zhou25y.html.

Fei Zhu, Xu-Yao Zhang, Zhen Cheng, and Cheng-Lin Liu. Revisiting confidence estimation: Towards reliable failure prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3370–3387, 2024. doi: 10.1109/TPAMI.2023.3342285. URL https://arxiv.org/abs/2403.02886.

A Proofs of Interface Bounds

We apply standard two-point and partial-identification arguments (Le Cam, 1986; Manski, 2003) to the benchmark observation interface.

Proof of Proposition 2. Fix an assignment on $|S| = k < 4$ edges. Among its 2^{4-k} completions, exactly half have either parity. The uniform component contributes $(1-\lambda)/2^k$, while the selected-parity component contributes $(\lambda/8)2^{3-k} = \lambda/2^k$. Their sum is 2^{-k} for either sign.

On the eight even states, the even and odd laws assign masses $(1+\lambda)/16$ and $(1-\lambda)/16$, respectively, with the roles reversed on the eight odd states. Summing the two constant log-ratios yields the stated KL divergence. □

Known-group MCAR corollary. For one known four-edge parity group with per-edge independent missingness rate r, parity signs differ only when all four edges are observed. The per-world observed-data divergence is therefore

$$(1-r)^4 \lambda \log \frac{1+\lambda}{1-\lambda},$$

and W independent worlds multiply it by W. The full 16-edge task carries additional mechanism information in the other groups, which share the task's mechanism and strength.

Proof of Proposition 1. Because (O, U) has the same distribution under both tasks, so does T. Let $\bar{p} = (p_0 + p_1)/2$. Then

$$\mathbb{E}(T-p_0)^2 + \mathbb{E}(T-p_1)^2 = 2\mathbb{E}(T-\bar{p})^2 + \frac{\Delta^2}{2} \geq \frac{\Delta^2}{2},$$

so the larger of the two risks is at least $\Delta^2/4$.

A deterministic interval containing both targets must span their distance. For the randomized case, let $A_i = \{p_i \in I\}$. Since both events are evaluated under the same law of (O, U),

$$\Pr(A_0 \cap A_1) \geq 1 - 2\alpha.$$

On this event $|I| \geq \Delta$, giving the stated expected-width bound. □

Collision floors by enumeration. All targets in Table 1 were verified by rational enumeration of the 2^4 or 2^8 joint states of the involved groups at every strength, and all 24 within-group edge permutations reproduce the star/path pair. For a balanced pair with known mechanism and strength, the optimal common prediction is the midpoint, with risk $\Delta^2/4$. For the star/path pair, $\Delta = (1-\lambda)/16$ under common shock or even parity and $(1+\lambda)/16$ under odd parity; averaging over the five strengths gives $33/102400$ and $233/102400$. For the aligned and crossed unions under common shock, with $a = (1+\lambda)/4$, the targets are $2a - a^2$ and $1/2 - a^2$, so $\Delta = 2a - 1/2 = \lambda/2$ and the floor is $\mathbb{E}[\lambda^2]/16 = 33/1600$; under parity both targets equal $7/16$. The support floor averages $\Delta^2/4$ between the four-edge conjunction and the two-pair union over the three mechanisms and five strengths, giving $721/25600$.

No-world references. Without worlds, the optimal shape-only predictor on the support pair uses the constants $13/96$ and $43/96$, with conditional risks $239/14400$ and $17/14400$ and combined risk $2/225 = 0.00889$. On all four base templates the aligned union and the two-pair union have the same named-formula distribution after relabeling; their hidden-template means are $43/96$ and $789/1600$, and the no-world named-query risk is 0.0097901.

B Protocols, Cohorts, and Inference

Factor grid. Every cohort crosses four topologies, its query templates, five interaction strengths, three missingness rates, and three context sizes. With the four base templates this gives 720 cells per mechanism; with seven templates, 1,260. Latent dependency groups are sampled independently of visible topology before edge and node permutations. Common shock and even parity are the training families; odd parity supplies deployment-shift tasks.

Table 7: Seed namespaces. Data seeds are disjoint across partitions and cohorts; model seeds control initialization and minibatch order only.

Cohort	Partition	Tasks	Data seeds
A	Train (10 runs)	$10 \times 32{,}768$	32001–32010
A	Calibration ID / test ID / test OOD	1,440 / 2,880 / 1,440	33001 / 34001 / 35001
B, D	Train (10 runs)	$10 \times 32{,}768$	42001–42010
B	Test ID / test OOD	2,880 / 1,440	44001 / 45001
C	Train (5 runs)	$5 \times 32{,}768$	52001–52005
C	Test ID / test OOD	2,520 / 1,260	54001 / 55001
D	Star/path test ID / OOD	2,880 / 1,440	47001 / 48001
P, T	Train (5 runs)	$5 \times 32{,}768$	42001–42005
P, T	Test: as B; T also as D	–	as B, D

Cohorts. Cohort A is the frozen evaluation cohort with ten one-epoch runs per interface, used for topology controls, detection, and risk ranking; its protocol, including the choice of 32,768 training tasks per run, was fixed on separate development seeds before final evaluation. Cohort B shares test tasks and ten data and model seed pairs across two sets of runs. The one-epoch shape diagnostic trains Set-mask, Set-shape, and Set-full and gives the support contrasts in Table 4. The budget sweep retrains Set-shape and Set-full for sixteen epochs (width 64: eight), saving predictions after epochs 1, 2, 4, 8, and 16. The sweep draws batches within world-count buckets without padding, which is exact because padded worlds are masked; with a constant learning rate and seeded per-epoch permutations, the checkpoint at epoch e equals an e-epoch run. Cohort C adds the star, path, and crossed templates to training and evaluation, trains Set-mask, Set-shape, and Set-full for five runs, and was specified, including seeds, estimands, and budgets, before its final phase was run. Cohort D uses the same sweep, trains on the base templates plus star and path, and is evaluated on a separate star/path test set. Cohorts P and T repeat the first five runs of B and D with the same data and model seeds for eight epochs, changing only the training families: P trains on even and odd parity with the base templates, and T on all three families with the templates of D. The moment and structure MLPs are retrained on the base templates under the same two priors.

Crossed inference. All runs of a cohort evaluate the same tasks. Each bootstrap replicate resamples training-run indices and one shared task-index vector, stratified by mechanism, and recomputes the ensemble prediction, task losses, and any induced ranking. Paired interface contrasts resample the runs of both interfaces jointly because they share data and model seeds. Contrasts against a deterministic reference resample tasks only. We use 2,000 percentile draws.

Tie-expected AURC. Exact score ties are handled analytically. For a tie block of size m with loss sum S_B, preceded by n_0 admitted tasks with loss sum S_0, the expected AURC contribution under a uniform ordering within the block is

$$\frac{1}{N} \sum_{r=1}^{m} \frac{S_0 + (r/m) S_B}{n_0 + r}.$$

Each bootstrap replicate reconstructs tie blocks after resampling.

Fixed-grid sensitivity. Within each of the 720 factorial cells of cohort A, the paired-block bootstrap resamples the two replicate blocks; each block contains its paired common and even tasks and the odd task matched by cell and replicate. Factor weights and mechanism ratios stay fixed while within-cell replication varies. Both task populations use each draw, which preserves their overlap for the population contrast in Section 5.4.

C Estimator Definitions

C.1 Partition BMA, oracles, shape and incidence references, and GOF

Let S contain the query edges, $k = |S|$, and let candidate $a = (\pi, f, \lambda)$ range over every unlabeled partition π of S, family f, and strength λ on a grid. The supported reference uses $f \in \{\mathrm{common}, \mathrm{even}\}$

and $\lambda \in \{0, 0.1, \ldots, 0.9\}$, giving $5 \cdot 2 \cdot 10$ candidates for $k = 3$ and $15 \cdot 2 \cdot 10$ for $k = 4$. The all-family reference adds odd parity; its off-grid variant uses the 40 midpoints $\{0.0125, 0.0375, \ldots, 0.9875\}$, which exclude every true strength. For each partition block, the four-bit family law is marginalized onto the block, and the blocks multiply to a normalized $P_a(x_S)$. With observed coordinates M_w in world w,

$$L_a = \sum_{w=1}^{W} \log\Big[\sum_{x_S :\, x_{M_w} = x_{M_w}^{\mathrm{obs}}} P_a(x_S) \Big], \qquad \omega_a \propto \exp(L_a),$$

$$\hat{p}_{\mathrm{BMA}}(q) = \sum_a \omega_a \sum_{x_S} q(x_S) P_a(x_S).$$

The prior is uniform over candidate triples. Missing coordinates are summed out, and a completely unobserved world contributes zero. The query-group oracle replaces the partition with the true latent groups, uses all four edges of each group that the query touches, and averages over all three families and the supported grid; it is a privileged reference. The full-information oracle uses the true groups and positions of all four groups, multiplies their observed-data likelihoods, and averages the exact target over the benchmark prior, uniform over the three families and $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; a parity-prior variant restricts the families to even and odd parity. Its posterior probability of odd parity serves as an all-edge detection score. A parity-prior variant of partition BMA and of the query-group oracle restricts the families in the same way.

On crossed-pairing tasks two further exact references read only the query edges. The shape Bayes predictor averages the target over the three templates with shape $\{2, 2\}$ and support size four (two-pair, aligned, and crossed unions), every split of the support into latent blocks, all three families, and the true strength grid, weighting each candidate by its observed-data likelihood. It is a function of the shape interface, so its risk upper-bounds $R^*(Z_{\mathrm{shape}})$. The incidence Bayes predictor does the same with the named pairing known. Partition GOF is the profile score $c_{\mathrm{GOF}} = \max_a L_a / \max(\sum_w |M_w|, 1)$. For the zero-strength candidate every observed bit contributes $\log(1/2)$, so GOF has an exact plateau at $-\log 2$; values within 10^{-12} of it are mapped to $-\log 2$ so that numerically equal scores tie exactly.

C.2 Reliability scores and intervals

Confidence-oriented scores are c_{GOF}, the fourth moment $c_4 = \hat{\mu}_4$, ensemble confidence $c_{\mathrm{ens}} = -\operatorname{Var}(\hat{p}_q)$ across runs, and interval confidence $c_{\mathrm{width}} = -(U - L)$. Detection uses their negatives with odd parity as the positive class; AURC admits larger confidence first. The fourth moment is the mean of $\prod_{e \in S}(1 - 2X_e)$ over worlds that observe all four query edges; it is zero when $|S| \neq 4$ or when no world observes the full support, which produces genuine ties. The conservative interval takes the lower and upper monotone completions of the query in each world, averages them, and widens the result by $\sqrt{\log(2/\alpha)/(2W)}$ at $\alpha = 0.05$.

C.3 Networks and moment learners

Table 8: Shared Set configuration. Set-mask, Set-shape, and Set-full differ only in the query encoder; Incidence models replace the local update with message passing over visible adjacency.

Item	Value
Input	$[B, W, 16]$ worlds, observed mask, valid-world mask
Edge embedding	width $d \in \{32, 64\}$; relation, provenance, query role, and state parts
Edge update	2 residual layers, GELU and LayerNorm
Clause and query encoding	size and count embeddings, nonlinear pooling
World Transformer	1 layer, 4 heads, feed-forward width $4d$, no positional encoding
Prediction	learned query token, LayerNorm, linear head, sigmoid
Parameters	27,081 ($d = 32$); 105,361 ($d = 64$)
Optimizer	AdamW, learning rate 3×10^{-4}, weight decay 0.01, constant schedule
Training	batch 64; 32,768 tasks per epoch; 1–16 epochs
Loss	binary cross-entropy against exact soft targets

Edge state is 0 for missing, 1 for observed zero, and 2 for observed one, embedded separately in each world. For clause c, the query encoder sums that world's incident edge embeddings divided by $\sqrt{|c|}$, adds a learned $\log(1 + |c|)$ embedding, and applies a nonlinear update; it then sums clauses divided by the square root of their count, adds a log-count embedding, and applies another update. Context and query vectors form one token per world before the Transformer aggregates across worlds, so simultaneous edge states are preserved. Shape replaces named membership with each sorted clause size spread uniformly over the support; mask uses one support clause.

The moment MLP receives relabeling-invariant structure features (support size, clause count, clause-size counts, maximum edge degree, co-occurring-pair fraction, and the independent-edge value) and world features (inverse world count, observed fraction, empirical moments of orders one to four on the query support, availability of the fourth moment, and the fraction of complete rows). The structure MLP keeps only the structure features. Both are two-layer GELU networks of width 64, trained with AdamW at learning rate 10^{-3} for 40 epochs at batch size 256 on the same training tasks and seeds as the Set models. They are trained under each of the three priors of Section 5.3: common and even parity; all three families; even and odd parity.

D Additional Results

D.1 Budget curves by cohort

Tables 9 and 10 give the ensemble risks behind Figure 2. In cohort D the risks are on the star/path test set.

D.2 Statistical references on the base templates

Table 11: Statistical references on cohort B (benchmark prior), and the one-epoch Set-full ensemble of cohort A.

Predictor	Common	Even	Odd	Deployment
Independent edges (no worlds)	0.036634	0.000645	0.000645	0.012641
Partition BMA, supported families	0.005673	0.000576	0.001080	0.002443
Partition BMA, all families	0.006300	0.000524	0.000618	0.002481
Partition BMA, all families, off-grid	0.006147	0.000533	0.000634	0.002438
Query-group oracle	0.002066	0.000409	0.000446	0.000974
Full-information oracle	0.000642	0.000112	0.000120	0.000292
Moment MLP	0.003278	0.001271	0.002339	0.002296
Set-full, cohort A, one epoch	0.006568	0.001893	0.003483	0.003981

Adding odd parity to the BMA families lowers odd-parity risk from 0.001080 to 0.000618 and slightly raises common-shock risk, and excluding the true strengths from the grid changes little. On the star/path test set, all-family BMA reaches 0.000375 on the training families and the query-group oracle 0.000113.

D.3 Full-information oracle and crossed-pairing references

Table 12: Full-information oracle risk by cohort under the benchmark prior. Cohort C rows cover the aligned and crossed unions only.

Cohort	Common	Even	Odd	Parity pool	Deployment
A	0.000614	0.000133	0.000098	0.000115	0.000281
B	0.000642	0.000112	0.000120	0.000116	0.000292
C (aligned, crossed)	0.000295	0.000040	0.000011	0.000026	0.000116

Table 9: Ensemble probability MSE by training budget, cohorts B and D (ten runs).

Cohort	Model	Epochs	Common	Even	Odd	Parity pool	Deployment
B	Set-shape	1	0.00743	0.00168	0.00233	0.00200	0.00381
B	Set-shape	2	0.00555	0.00135	0.00223	0.00179	0.00304
B	Set-shape	4	0.00442	0.00104	0.00241	0.00173	0.00262
B	Set-shape	8	0.00392	0.00085	0.00202	0.00144	0.00227
B	Set-shape	16	0.00339	0.00089	0.00197	0.00143	0.00209
B	Set-full	1	0.00581	0.00204	0.00296	0.00250	0.00360
B	Set-full	2	0.00461	0.00145	0.00253	0.00199	0.00286
B	Set-full	4	0.00305	0.00107	0.00248	0.00177	0.00220
B	Set-full	8	0.00284	0.00082	0.00197	0.00140	0.00188
B	Set-full	16	0.00236	0.00091	0.00196	0.00144	0.00175
B	Set-full, width 64	1	0.00509	0.00168	0.00278	0.00223	0.00318
B	Set-full, width 64	2	0.00348	0.00135	0.00247	0.00191	0.00244
B	Set-full, width 64	4	0.00279	0.00098	0.00238	0.00168	0.00205
B	Set-full, width 64	8	0.00265	0.00080	0.00199	0.00140	0.00181
D	Set-shape	1	0.00055	0.00047	0.00385	0.00216	0.00162
D	Set-shape	2	0.00066	0.00058	0.00292	0.00175	0.00139
D	Set-shape	4	0.00061	0.00058	0.00283	0.00170	0.00134
D	Set-shape	8	0.00047	0.00044	0.00306	0.00175	0.00132
D	Set-shape	16	0.00039	0.00042	0.00310	0.00176	0.00130
D	Set-full	1	0.00082	0.00057	0.00368	0.00212	0.00169
D	Set-full	2	0.00077	0.00063	0.00274	0.00169	0.00138
D	Set-full	4	0.00045	0.00048	0.00253	0.00151	0.00115
D	Set-full	8	0.00033	0.00035	0.00275	0.00155	0.00114
D	Set-full	16	0.00030	0.00031	0.00235	0.00133	0.00099
D	Set-full, width 64	1	0.00111	0.00054	0.00367	0.00210	0.00177
D	Set-full, width 64	2	0.00079	0.00067	0.00250	0.00159	0.00132
D	Set-full, width 64	4	0.00035	0.00038	0.00260	0.00149	0.00111
D	Set-full, width 64	8	0.00028	0.00027	0.00252	0.00140	0.00103

Table 10: Ensemble probability MSE by training budget, cohort C (five runs, seven templates). Collision columns restrict to the two templates of each pair; all columns except "Crossed (common)" weight the three mechanisms equally.

Model	Epochs	All templates	Support pair	Star/path	Crossed (common)	Crossed (all)
Set-mask	1	0.01942	0.04559	0.00621	0.02659	0.00968
Set-mask	2	0.01781	0.04731	0.00473	0.02279	0.00836
Set-mask	4	0.01744	0.04641	0.00499	0.02294	0.00851
Set-mask	8	0.01727	0.04492	0.00564	0.02354	0.00900
Set-mask	16	0.01725	0.04517	0.00549	0.02338	0.00892
Set-shape	1	0.00434	0.00445	0.00127	0.02159	0.00730
Set-shape	2	0.00395	0.00341	0.00135	0.02088	0.00729
Set-shape	4	0.00328	0.00206	0.00128	0.02107	0.00714
Set-shape	8	0.00324	0.00195	0.00127	0.02124	0.00716
Set-shape	16	0.00318	0.00185	0.00134	0.02102	0.00709
Set-full	1	0.00371	0.00374	0.00122	0.01766	0.00608
Set-full	2	0.00327	0.00330	0.00133	0.01458	0.00536
Set-full	4	0.00219	0.00265	0.00109	0.00744	0.00288
Set-full	8	0.00181	0.00239	0.00109	0.00508	0.00192
Set-full	16	0.00161	0.00202	0.00102	0.00483	0.00179

Table 13: References and trained risk on the crossed-pairing collision, cohort C, common shock, aligned and crossed templates. The first block brackets $R^*(Z_{\mathrm{shape}})$.

Interface	Predictor	Probability MSE
Shape	Floor, mechanism and strength known	0.02063
Shape	Shape Bayes on query edges (upper bound)	0.02113
Shape	Set-shape, 16 epochs, ensemble / single	0.02103 / 0.02104
Incidence	Partition BMA, all families	0.00613
Incidence	Incidence Bayes on query edges	0.00473
Incidence	Set-full, 16 epochs, ensemble / single	0.00483 / 0.00509
True groups	Query-group oracle	0.00083
True groups	Full-information oracle	0.00030

D.4 STAR/PATH COLLISION BY TRAINING PRIOR

Table 14: Star/path test set (cohorts D and T), five- or ten-run ensembles. The shape floor binds every shape-only predictor with mechanism and strength known. Contrasts are paired with crossed run–task 95% intervals.

Prior	Model, epochs	Mechanisms	Set-shape	Set-full	Floor	Set-full minus floor [95% CI]
common, even	width 32, 16	training	0.000404	0.000305	0.000322	-0.000017 [-0.000041, $+0.000019$]
common, even	width 64, 8	training	–	0.000278	0.000322	-0.000044 [-0.000056, -0.000022]
common, even	width 32, 16	odd	0.003104	0.002351	0.002275	above floor
all three	width 32, 8	training	0.000646	0.000496	0.000322	$+0.000173$ [$+0.000138$, $+0.000258$]
all three	width 32, 8	odd	0.002551	0.002246	0.002275	-0.000029 [-0.000186, $+0.000167$]
all three	width 32, 8	all	0.001281	0.001079	0.000973	$+0.000106$ [$+0.000060$, $+0.000177$]

Set-full minus Set-shape is -0.00075 ($[-0.00090, -0.00059]$) on odd parity under the two-family prior, and -0.00015 ($[-0.00024, -0.00004]$), -0.00030 ($[-0.00042, -0.00017]$), and -0.00020 ($[-0.00024, -0.00016]$) on the training families, odd parity, and all mechanisms under the three-family prior.

D.5 RELIABILITY SCORES

For partition GOF, ensemble-minus-GOF AURC on cohort A is -0.002594 on the deployment mixture (paired-block CI $[-0.004787, -0.001249]$) and -0.001340 on the even/odd pool ($[-0.002316, -0.000791]$); the population contrast is -0.001254 ($[-0.003560, +0.000040]$). The conservative interval covers 0.9991 of targets with mean width 0.652.

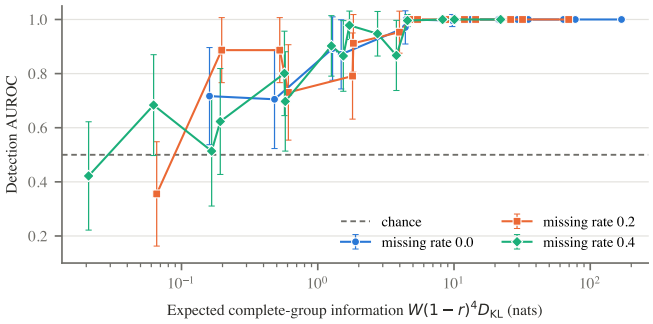


Figure 3: Fourth-moment odd-parity detection on cohort A against the expected complete-group information $W(1-r)^4 D_{\mathrm{KL}}$ of the query support. Each cell pools 16 even and 16 odd parity-sensitive tasks; bars are Hanley–McNeil 95% intervals. The three missingness levels fall on one curve.

D.6 SUPPORT FLOOR AND TOPOLOGY CONTROLS

Table 15: Analytic support-mask floor and one-epoch ensemble risk on the support pair, cohort A. Mechanism rows average the five strengths; the final row averages the mechanism–strength grid.

Quantity on the support pair	Probability MSE
Support floor, common shock	0.01160
Support floor, even parity	0.02473
Support floor, odd parity	0.04816
Support floor, mechanism–strength grid	0.02816
Set-mask ensemble	0.03847
Set-full ensemble	0.00380
Set-mask minus Set-full	0.03467

Table 16: Paired one-epoch probability-MSE differences on cohort A, crossed run–task bootstrap.

Paired comparison	Difference	95% CI
Set-full minus Set-mask, support pair	-0.0343	$[-0.0368, -0.0319]$
Incidence-full minus Set-full, all tasks	-0.0006	$[-0.0029, +0.0016]$
Incidence minus shuffled adjacency, all tasks	-0.000028	$[-0.000100, +0.000032]$

E Data, Artifacts, and Reproducibility

The benchmark is synthetic and requires no external dataset. The anonymous supplement contains the generator, configurations, training and analysis code, tests, and prediction-level outputs for all four cohorts: per-run neural predictions at every budget checkpoint, exact targets, statistical-reference predictions, detection scores, and factorial-cell metadata. These files reproduce every interval and tie-expected AURC analysis without retraining, and unit tests check each generated target against rational enumeration. SHA-256 manifests authenticate the released artifacts.