

# GRAPHS ARE EVIDENCE, NOT CIRCUITS

Aditya Gaur Charu Sharma

Machine Learning Lab, IIIT Hyderabad

## ABSTRACT

Graph learning has mistaken its evidence for its machinery. The evidence says what is observed; the query says what must be inferred; the circuit specifies how observations are combined. Message passing commonly inherits its circuit from the graph. We make that choice explicit. On identical evidence, two binary queries with the same prior and Bayes AUC require different interactions: a linear readout is optimal for one but stays 9.6 AUC points below it for the other, even after all observations are in reach. A product readout closes the gap. On Minesweeper, recompilation improves a depth- and parameter-budget-matched GCN by 4.28 AUC points; scrambling alignment while preserving the entire circuit topology reduces performance to near chance. **Two-System Graph Learning (TSGL)** puts the distinction to work: preserve a capable GNN, learn explicit relationships within a declared library, and deploy additional inference only after held-out acceptance. Without label inputs, TSGL reaches state-of-the-art-level Minesweeper performance: 99.13 AUC against a validation-selected GNN/ensemble baseline of 98.19, removing 52% of its remaining ranking error. With training-label anchors, TSGL reaches 99.97 AUC in a separate matched panel, removing 98% of the remaining ranking error. Gains beyond generic correction and independent GNN ensembles transfer from arxiv-year to predeclared views of a fresh patent graph. On three fresh social graphs, the typed library does not beat those controls; the system deploys them instead, improving 14 of 18 evaluations and exactly preserving the rest. In our tests, every accepted deployment improved the test score; every refusal returned the baseline exactly. A separate count-family compiler reuses learned relationships on independent boards and joint queries without target fitting. Treat the graph as evidence, then build the computation the question requires.

## 1 INTRODUCTION

Consider the two Minesweeper clues in [Figure 1](#). What do they tell us about the unseen cell  $x_1$ ? One says that  $x_1$  and  $x_2$  contain one mine; the other says that  $x_2$  and  $x_3$  contain two. The second clue fixes  $x_2 = x_3 = 1$ . Substituting into the first gives  $x_1 = 0$ . We have answered the question by composing two observations. The grid identifies neighbours. The deduction uses a different object: a sequence of operations on overlapping constraints.

Message-passing graph neural networks (GNNs) commonly use their input graph twice: once as the carrier of evidence, and once as the circuit that assembles an answer ([Kipf & Welling, 2017](#); [Hamilton et al., 2017](#); [Veličković et al., 2018](#)). This identifies two different objects. Observed edges describe relationships in the data; computational dependencies specify how to combine observations for a query. Their identification is an inductive bias, not a requirement of graph learning. Even a graph irrelevant to the target can influence a GNN ([Bechler-Speicher et al., 2024](#)).

The central claim is that *graph learning must distinguish the structure of observations from the structure of inference*. A graph can tell a learner which observations are related without specifying the operations needed to answer a particular question. More propagation, different local operations and different computational dependencies are therefore distinct resources. Comparing architectures on a fixed adjacency need not isolate their contributions.

We develop this argument through three contributions.

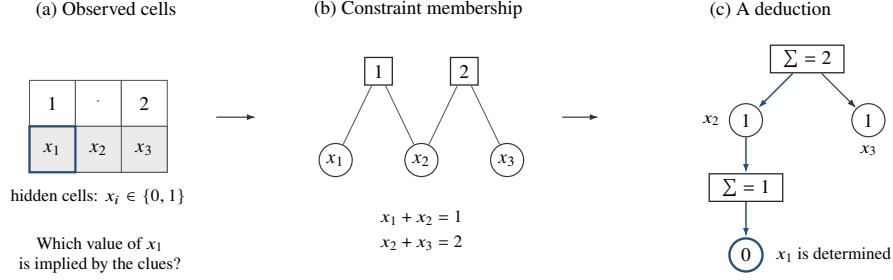


Figure 1: **The observation graph does not specify the deduction.** Panel (b) groups the variables in each clue; panel (c) composes the constraints to determine  $x_1$ . No new observation is added. Only the displayed clues are used; the known-empty dot’s clue is omitted.

**A formulation.** Evidence, query and inference circuit are separately variable parts of a graph-learning problem. The observed graph can be reused as a circuit, but that identification is an explicit modeling choice.

**A controlled analysis.** We intervene on circuit membership, local operations and anchor access. An isomorphic scramble preserves the entire typed topology while breaking its alignment with the evidence. Correct anchors help the aligned circuit but cost 19.2 AUC points under random membership: more correct evidence can amplify the wrong computation. Holding evidence, binary class prior and Bayes AUC fixed, one query admits an optimal linear ranking; the other requires an interaction that greater radius cannot supply. Fitted curves follow the exact ranking predictions.

**Two constructive consequences.** **Two-System Graph Learning (TSGL)** preserves a capable neural predictor and tests whether explicit relational computation improves it. A separate count-family compiler tests whether a learned relationship remains executable across independent instances and new joint queries. Together they test predictive supplementation and reusable inference.

Minesweeper is a flagship consequence: near-perfect prediction, recovery of exercised counting relationships, and reusable joint inference. Against a validation-selected GNN/ensemble baseline, strict TSGL raises AUC from 98.19 to 99.13, removing 52% of the remaining ranking error without label inputs. With training-label anchors in the ten-task panel, TSGL reaches 99.97 AUC and removes 98% of its matched baseline’s remaining ranking error. On arxiv-year and predeclared views of a new patent-citation graph, typed supplementation beats matched generic correction and independent-model ensembling. Fresh social-graph tests make the boundary equally concrete: generic or ensemble corrections improve the full system, while the tested typed library does not beat those controls.

The claim is not that every graph needs the same reasoning engine. Separating evidence from computation makes circuit design an explicit, testable axis of graph learning, with consequences for both prediction and reusable knowledge.

## 2 SEPARATE THE EVIDENCE, THE QUERY AND THE CIRCUIT

*In brief. Observed relationships and computational dependencies are different objects. We intervene on the latter without adding evidence.*

Let  $\mathcal{E} = (G, X, O, M)$  collect the observed evidence: graph, features, revealed values and observation masks. A query  $q$  specifies a target  $y_q$ ; an inference circuit  $\mathcal{C}_q$  specifies intermediate quantities, dependencies and update operations:

$$\underbrace{\hat{y}_q}_{\text{answer}} = \text{Execute} \left( \underbrace{\mathcal{C}_q}_{\text{circuit}} ; \underbrace{\mathcal{E}}_{\text{evidence}}, \underbrace{\theta}_{\text{parameters}} \right). \quad (1)$$

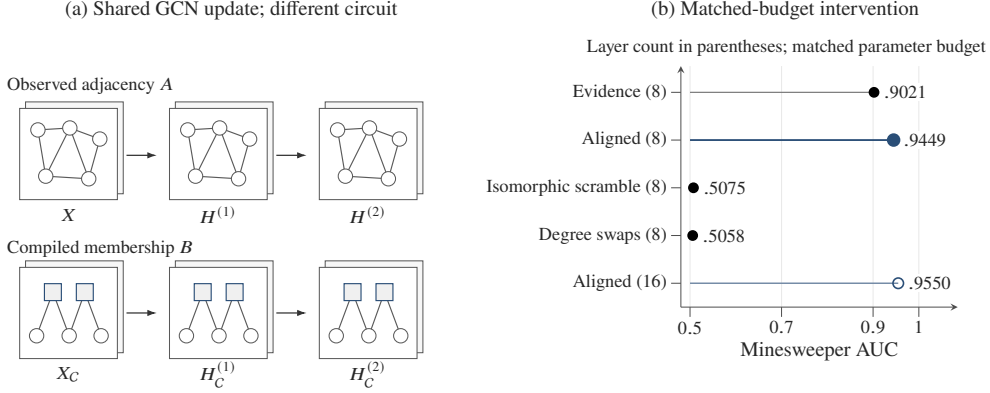


Figure 2: **Alignment, not graph shape alone, carries the circuit gain.** Left: a shared GCN update family on observed adjacency or compiled membership. Right: tie-aware AUC over five splits and five seeds (ten fits for degree swaps). The isomorphic scramble keeps the whole typed topology, features, architecture, depth and initialization, but changes which variables feed each factor. The eight-layer contrast is primary; sixteen compiled layers are a higher-depth reference.

A circuit may reuse adjacency, introduce factor nodes, or follow a trellis. Equation (1) makes that choice explicit rather than inheriting it from observed connectivity. The task is the tuple (*generator, observation model, query, training access*), not the graph alone. An *anchor* is a revealed target value that inference may use.

This formulation separates two questions. Does a computation combine the right observations in the right way? Can a learner identify it from its permitted training data? We first test circuit alignment, then isolate query-dependent interaction requirements in Section 3.

## 2.1 KEEP THE QUESTION; CHANGE THE CIRCUIT

On Minesweeper (Platonov et al., 2023), a revealed clue constrains the number of mines among its neighbours. We compile this supplied interpretation into a bipartite circuit: one factor per revealed clue, connected to its observed neighbours, with the clue value copied onto the factor. No hidden value or extra target label is introduced. Variables retain their original features; variable–factor incidences replace the original variable–variable edges.

All arms use the same residual GCN update family, training targets and optimization recipe. The primary comparison keeps eight layers and matches parameter counts within 0.7%. Across five official splits and five optimizer seeds, the compiled circuit reaches .9449 AUC against .9021 on evidence adjacency: **+4.28 AUC points, positive in all 25 pairs**. The bipartite input has separate variable/factor encoders. The following control holds those encoders and the complete circuit architecture fixed.

**A topology-preserving test of alignment.** Let  $B$  be the factor–variable incidence matrix. We construct  $B' = B\Pi$ , where  $\Pi$  permutes variables within equal-incidence-degree classes. Factor features and variable features, labels and masks stay at their original indices. Both degree sequences and the entire typed topology are preserved up to isomorphism, including cycles and factor overlaps. What changes is the alignment between computational membership and evidence. Architecture, width, depth, parameters and initialization match the aligned arm exactly.

The scrambled circuit scores .5075, versus .9449 with aligned membership: a 43.74-point difference, positive in all 25 pairs. Only .09–.24% of original incidences remain aligned. A co-permutation check, which moves features together with the graph, preserves predictions to numerical precision. Degree-preserving bipartite edge swaps also score near chance (.5058). Thus the aligned gain cannot be attributed solely to an extra node type, parameter allocation, degree sequence or unlabelled topology.

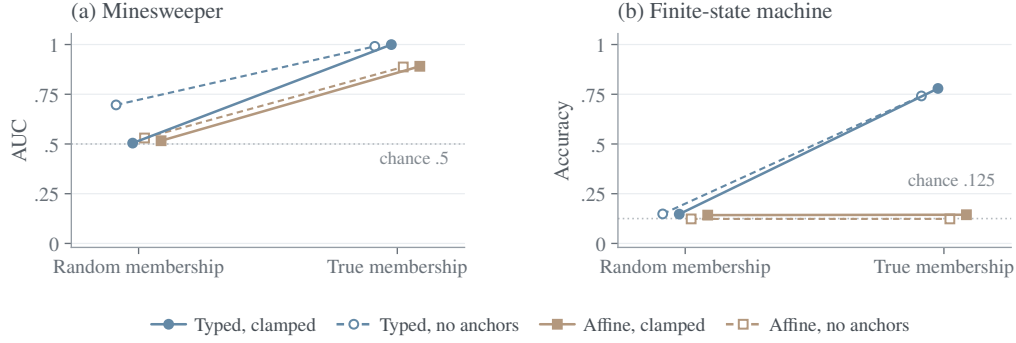


Figure 3: **An operation is useful in relation to its membership.** Blue circles: typed kernels; sand-colored squares: affine kernels. Solid lines and filled markers clamp anchors; dashed lines and hollow markers omit them. Dotted lines mark chance. Minesweeper averages 25 split/seed fits per cell; FSM averages five draws. All means and protocols: [Appendix E.2](#).

Sixteen compiled layers reach .9550, a higher-depth reference rather than an update-matched comparison: traversing two bipartite edges does not equate nonlinear operations or receptive fields. Recipes, paired results and spatial-resampling sensitivity are in [Appendix E.1](#).

The value of recompilation is task-dependent. In a descriptive Cora control, replacing evidence edges with a neighbourhood-factor scaffold changes accuracy by  $-3.3$  points; augmenting the original graph changes it by  $+.25$ . That scaffold carries no supplied clue semantics. The contrast supports choosing computational dependencies for the task, not inserting factor nodes indiscriminately ([Appendix E.1](#)).

## 2.2 MEMBERSHIP, OPERATORS AND ANCHORS INTERACT

A  $2 \times 2 \times 2$  design crosses true/random membership, typed/affine kernels and clamped/absent anchors. Minesweeper uses five official splits and five optimizer seeds. With anchors, typed kernels add 10.91 AUC points under true membership, positive in every pair, but  $-1.20$  under random membership. A finite-state-machine (FSM) factorial, evaluated with a fixed belief-propagation solver on five draws, gives corresponding gains of 63.49 and .49 accuracy points. Useful local operations depend on connecting the appropriate variables.

**Correct observations can amplify a wrong circuit.** Under random typed membership, clamping training anchors reduces Minesweeper AUC from .6964 to .5043, a 19.21-point drop. Under true typed membership, the same intervention raises AUC from .9909 to .9997. Every split/seed pair has the corresponding sign. Anchors inject evidence into a hypothesis about dependencies; their correctness does not repair an incorrect hypothesis. The direction of this effect also holds in the FSM factorial. Frozen predictions and outcomes are retained in [Appendix F](#).

## 3 REACH IS NOT INTERACTION: AN EXACT SEPARATION

*In brief. Radius changes which observations a linear score reaches, not the algebraic interactions it can express.*

Let observation signs be  $\sigma \in \{-1, +1\}^K$ , with affine encoding  $X(\sigma)$ . For a linear propagation operator  $P$  fixed with respect to the evaluated observation, set

$$F_r = [X, PX, \dots, P^r X], \quad S_i(\sigma) = w^\top F_r[i] + \beta. \quad (2)$$

**Theorem 1** (Affine confinement). *Every score in Equation (2) is affine in  $\sigma$ , at every radius. Concatenating multiple fixed linear operators or fixed affine preprocessing does not change this conclusion.*

Linear maps preserve affine functions: increasing radius or changing fixed connectivity cannot introduce a missing product into this score class. A fixed strictly increasing logistic link preserves its binary ranking. The decisive object is the posterior *ordering*, not merely a nonlinear posterior; the next result gives sufficient conditions for a strict AUC separation.

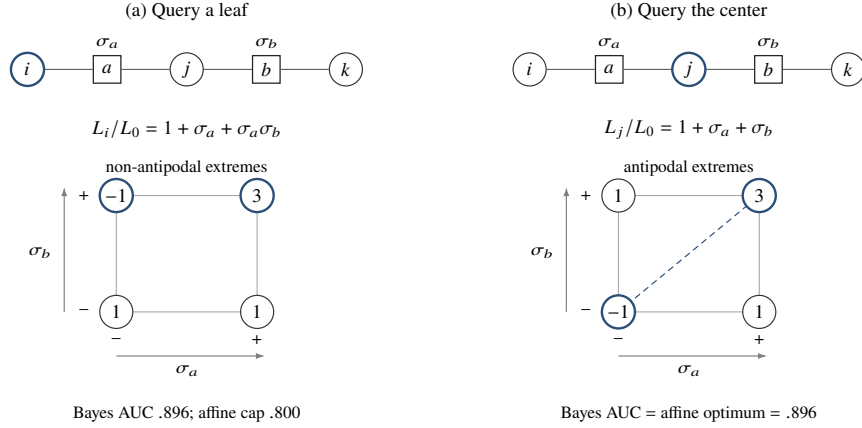


Figure 4: **Same evidence and Bayes difficulty, different ranking requirements.** The query changes from the leaf to the center of the same two-check construction. Numbers at the observation-square corners are posterior log-odds divided by  $L_0$ , in favor of zero. The leaf’s unique extreme atoms are not antipodal; the center’s log-odds are affine. Numerical AUCs use  $p = .2$ .

**Theorem 2** (Sufficient ranking obstructions). *On a finite sign cube with positive class-conditional atom masses, the best affine AUC is strictly below Bayes if either (a) the likelihood ratio has unique highest and lowest atoms that are not antipodes, or (b) both class-conditional distributions are sign-flip invariant and their likelihood ratio is nonconstant. In (b), every affine score has AUC  $1/2$ .*

An affine score’s unique extrema are antipodal. Under sign-flip symmetry, its class-conditional scores are symmetric about the same center. These observations give the ranking obstructions; finiteness makes the gap strict uniformly over affine orderings (Appendix A).

Table 1: **The query determines the affine gap on the same evidence.** Rows 1–2 use independent Bernoulli( $p$ ) bits and  $0 < p < 1/2$ . The balanced noisy-parity model in row 3 is defined in Appendix A, with  $0 < \varepsilon < 1/2$ . Entries are population-optimal AUCs in the stated classes.

| Construction / query | Bayes AUC                       | Best affine AUC | Gap                    |
|----------------------|---------------------------------|-----------------|------------------------|
| Two checks / leaf    | $(1-p)^2(1+2p)$                 | $1-p$           | $p(1-p)(1-2p)$         |
| Same checks / center | $(1-p)^2(1+2p)$                 | Same as Bayes   | 0                      |
| Noisy parity         | $1-2\varepsilon(1-\varepsilon)$ | $1/2$           | $2(\varepsilon-1/2)^2$ |

For independent bits  $i, j, k$ , observe  $\sigma_a = (-1)^{i+j}$  and  $\sigma_b = (-1)^{j+k}$ . With  $L_0 = \log((1-p)/p)$ , the posterior log-odds in favor of zero are  $L_i = L_0(1 + \sigma_a + \sigma_a \sigma_b)$  for the leaf and  $L_j = L_0(1 + \sigma_a + \sigma_b)$  for the center. The same evidence gives two queries with the same binary prior and the same Bayes AUC, but different affine gaps (Figure 4). At  $p = .2$ , Bayes AUC is .896 for both; the leaf’s best affine score is capped at .800, whereas the center admits an optimal affine ranking.

### 3.1 KEEP THE EVIDENCE; CHANGE THE QUESTION

We realize the construction as a graph of 20,000 disjoint paths  $i - a - j - b - k$ . Check nodes carry the signs; variable nodes carry only type indicators. Whole gadgets are assigned to training, validation and test, and the evidence array and gadget split are identical for the two target fits. The readout is taken at  $i$  or  $j$ . Propagation is fixed row-normalized adjacency with self-loops, with no data-dependent feature normalization. Linear logistic, quadratic-feature logistic and one-hidden-layer MLP readouts use the same radius-indexed feature bank and validation-only regularization selection.

For the leaf, the first check is reachable at radius one, but the second first arrives at radius three. At radii one and two, all three fitted readouts attain .800 population AUC. From radius three onward, the product and MLP readouts reach .896, while the linear score remains at .800 through radius six. For the center, linear readout attains .896 from radius one. At radius zero, every readout is at

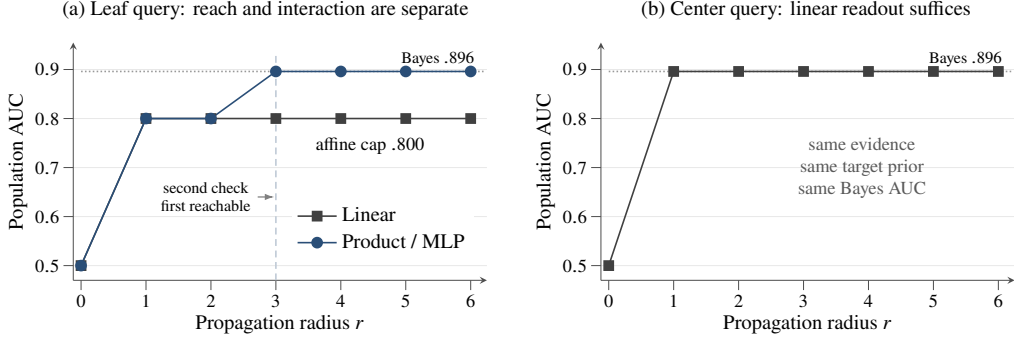


Figure 5: **Seeing the second check helps only a readout that can combine it.** Fitted population AUC at  $p = .2$ , five independent gadget draws. Left: the distant sign first reaches the leaf at radius three; product and MLP readouts then attain Bayes, while the linear score stays at its exact cap. Right: the center needs no product and its linear readout reaches the same Bayes AUC at radius one. Curves coincide across draws at the displayed precision.

chance. These are the fitted models’ population AUCs, computed by evaluating their scores on the four observation atoms with exact class-conditional weights; held-out gadget scores are also recorded.

The pattern holds at  $p = .05, .10, .30$ , with the corresponding exact optima. On balanced noisy parity with corruption .10, fitted affine scores give AUC .500 and product/MLP readouts give .820. The proof specifies a function-class obstruction; the experiment checks that fitted learners attain the predicted rankings. Together they separate two failure modes that a radius sweep alone would conflate: an observation has not arrived, or the readout cannot express the required interaction after it arrives. Full fitting and scoring details are in [Appendix A.7](#).

**From expressivity to learning.** The separation concerns fixed linear propagation and affine readout. Nonlinear GNNs, transformers and observation-dependent operators can create richer interactions ([Xu et al., 2019](#); [Rampášek et al., 2022](#)); learning the useful ones under a given access budget is an empirical question. The theorem identifies a resource that reach alone cannot provide in its score class; the GCN study tests explicit circuit construction within a declared nonlinear-model budget.

## 4 THE QUESTION CHANGES THE LEARNING PROBLEM

*In brief. A useful diagnostic measures what a stated readout can learn about a stated query from the available evidence.*

The matched binary queries isolate an algebraic requirement. A finite-state task tests the same design principle with partial observations and a numerical law that the learner is not given. A state machine evolves as  $s_{t+1} = \delta(s_t, x_t)$ . Its 20,000 time positions form an undirected path. Each node carries a noisy symbol observation; a shared 2% observation mask reveals both the true state and symbol. We construct one evidence tensor and ask either for a hidden state or for a true symbol.

Observation anchors and supervision occupy disjoint positions. A set  $S$ , of the same size as the anchor set  $O$ , supplies training targets but never label inputs. Radius and weight decay are selected on a separate validation set; both questions use identical observation, supervision, validation and test masks. A forward-backward reference receives the true transition table and noise law, together with the same observation anchors. Its privilege is knowledge of the numerical law, not additional observed test values.

Across five draws in two state-space configurations, the reference exceeds the validation-selected linear state probe by 49.04 accuracy points on average (range 47.8–51.0). The symbol probe is within .05 point on average ([Figure 6](#)). A validation-selected GCN also leaves a large state deficit. This is a latent-state versus near-observed-symbol contrast: corruption is 5%, so the symbol question has a strong local observation. Equal Bayes difficulty is established by the binary construction, not assumed for this FSM comparison.

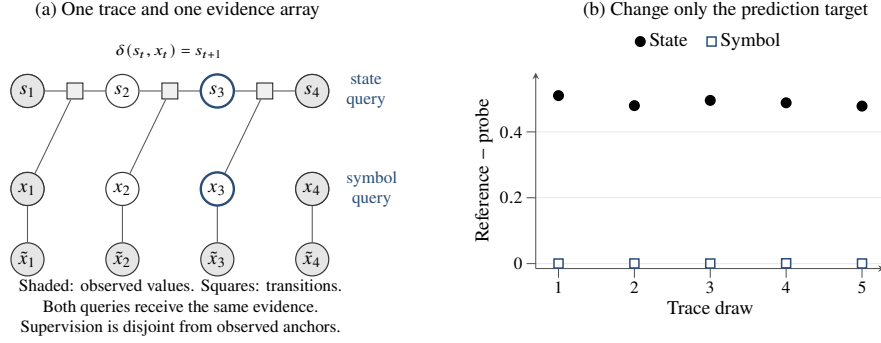


Figure 6: **One evidence array, two learning problems.** Both queries receive all shaded observations. Supervision is disjoint from those anchors, and the readout radius is validation-selected. Exact-reference minus probe accuracy remains large for states and near zero for symbols. Draws 1–4 have eight states and four symbols; draw 5 has sixteen states. Zero is equality to the reference, not chance accuracy. Protocol and all draw-level scores: [Appendix E.3](#).

**Readout-specific headroom, not a universal ceiling.** Linear probes measure which evidence is accessible to an affine readout; nonlinear heads test additional interactions; a law-aligned reference measures performance attainable with the supplied numerical model. These comparisons distinguish missing reach, missing interaction and learning access. They do not turn a low probe score into an impossibility theorem for nonlinear learners. Empirical diagnostic panels, including their scope and unsuccessful predictions, are retained in [Appendices D and F](#). TSGL selects candidates empirically rather than treating a probe reading as an automatic deployment decision.

## 5 TWO-SYSTEM GRAPH LEARNING

*In brief. Preserve strong perception and add the computation that improves it.*

TSGL separates perception, structured inference and acceptance ([Figure 7](#)). Validation nodes are divided into a selection set  $V_{\text{sel}}$  and an acceptance set  $V_{\text{acc}}$ . The system retains a saved baseline  $p_0$ , fits specialist candidates on permitted raw evidence, and selects a probability mixture on  $V_{\text{sel}}$ :

$$p_\lambda = (1 - \lambda)p_0 + \lambda p_s, \quad \lambda \in \{.25, .5, .75, 1\}. \quad (3)$$

The design objective is *complementarity*: learn corrections that improve  $p_0$ , with the preserved predictor itself a first-class execution path. Without acceptance, TSGL returns  $p_0$  pointwise. Both access protocols fit training targets: *strict* forbids label inputs; *anchored* also permits training labels as inputs.

**Perception and evidence.** The perceiver is selected from declared GCN, GraphSAGE, GAT, APPNP or directed-SAGE recipes, with the menu stated per evaluation. The strong Minesweeper comparison also permits an equal-weight ensemble of three reproduced GNN families. Specialists receive original and fixed propagated features, preserving incoming/outgoing roles on directed graphs. Raw evidence remains accessible; baseline probabilities enter the output mixture directly, rather than being reconstructed by another head. Training, checkpoint/weight selection, acceptance and test nodes are disjoint within each fit. [Appendix B](#) gives masks and menus.

**The relational branch.** For at most ten classes, the engine uses learned affine, cardinality and parity updates, ten damped rounds, and feature-affinity membership capped at eight neighbours per role. One membership policy was chosen on development tasks and frozen; ties are broken by node ID, making membership invariant to edge-storage order. All-edge feature propagation remains available; explicit higher-order factors use the retained members. Larger class counts use an uncapped pairwise approximation.

The typed library specifies *how a relationship can depend on its members*; training fits its numerical tables and type weights. A count table indexes how many neighbours occupy a class, rather than

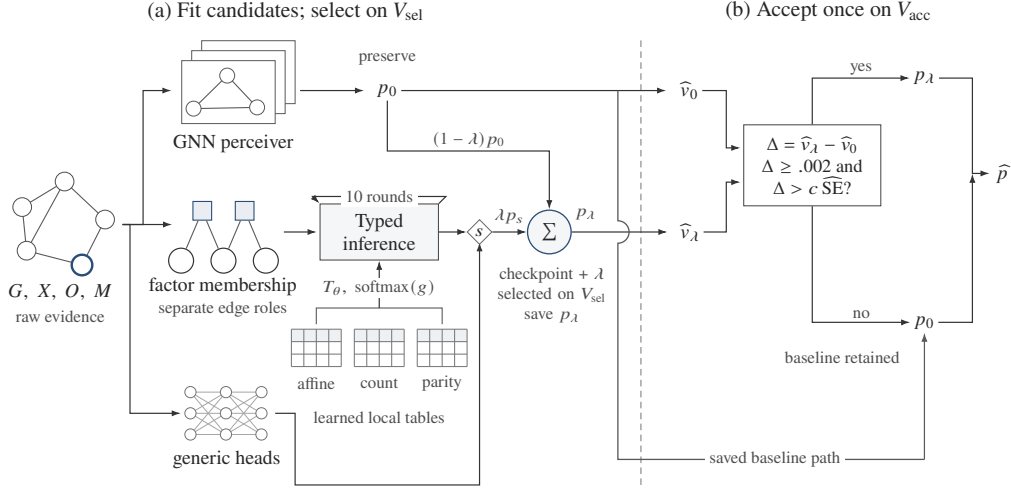


Figure 7: **Preserved perception with learned relational supplementation.** Fit tables  $T_\theta$  and type gates  $g$  on training targets; choose the checkpoint, candidate and mixture weight on  $V_{\text{sel}}$ . Lock the proposal before  $V_{\text{acc}}$ . Deploy the accepted mixture or retain the saved baseline pointwise, without executing the refused specialist. An independently fitted GNN supplies an ensemble candidate under the same selection/acceptance interface (not drawn).

assigning an unrelated parameter to every joint configuration. A parity table indexes the count modulo two; the affine branch restricts the dependence to a linear count term. The same learned objects are reused during successive message updates. Local contractions are explicit, but the full loopy multiclass recurrence is approximate. [Appendix B](#) gives the table parameterization and updates.

Two generic correction heads receive the same feature bank and comparable selection access. An independently trained second predictor supplies an ensemble control. Each candidate is selected for its contribution to the retained baseline, not for standalone performance. These controls test whether the structured branch contributes more than additional fitting or another predictor.

**Isolated acceptance and exact preservation.** A saved proposal is accepted only when its acceptance gain is at least .002 and exceeds  $c$  paired standard errors:  $c = 2.24$  in the ten-task panel and 2.50 in the strong-baseline, arxiv-year and fresh-data comparisons. Accuracy uses paired correctness differences; AUC uses paired positive/negative placement differences. The preservation identity is exact, without a sampling assumption. Acceptance on dependent graph observations is an empirical decision rule, not a guarantee of future no-harm. [Appendix B](#) gives the computation; [Appendix A](#) distinguishes the older IID argmax analysis.

Table 2: **Separate the learning problem from the information supplied.** Training targets, label observations and structural assumptions are distinct resources. Each branch below has its own explicit access contract.

| Branch                        | Target-label fitting | Label inputs at inference | Supplied structure                     |
|-------------------------------|----------------------|---------------------------|----------------------------------------|
| Integrated, strict            | Yes                  | None                      | Graph; engine’s typed library/policy   |
| Integrated, anchored          | Yes                  | Training labels only      | Same; access-matched baseline          |
| Original CFN, no label inputs | Yes                  | None                      | Cardinality family and clue membership |
| Frozen-table reuse            | Source only          | Declared target anchors   | Count family, grid and local compiler  |
| Supplied-law reference        | No gradient fitting  | Declared anchors          | Family, membership and numerical law   |

The anchored baseline is selection-best among the original predictor, a label-input adaptation and Correct&Smooth (Huang et al., 2021). Label-input fitting masks a random half of training labels as observations and predicts the remainder; acceptance/test labels never become inputs. Anchored gains are therefore measured against a baseline that can use the same label observations.

## 6 CONSTRUCTIVE CONSEQUENCES: PREDICTION AND LEARNED LAWS

*In brief. Typed computation improves strong predictors on Minesweeper and citation-time tasks; matched alternatives establish where it adds value.*

Across the strong Minesweeper comparison, the ten-task panel, the arxiv-year controls, the patent views and the social panel, every accepted full-system proposal improves test performance relative to its retained baseline, and every refusal returns that baseline pointwise. This is measured selective supplementation: gains when a proposal earns deployment, exact preservation otherwise. These evaluations share graphs, splits and some fitted proposals, so we report their accounting separately rather than pool them as independent successes (Appendix B.1).

The circuit and factorial interventions, strong-baseline comparison and fresh-data evaluations use preregistered endpoints and selection rules. Frozen predictions, scored misses, exposure chronology and exploratory controls are distinguished in Appendix F and the corresponding protocols.

### 6.1 MINESWEEPER: IMPROVE STRONG PERCEPTION AND RECOVER THE RELATIONSHIP

The strict-access comparison preserves the best validation-selected predictor from reproduced GCN\*, GraphSAGE\* and GAT\* recipes (Luo et al., 2024), or their equal-weight ensemble. All candidates use the same official masks, with validation divided into selection and acceptance before fitting. Across nine confirmation splits, the retained predictor scores 98.189 AUC; a mixture with an independently trained second predictor scores 98.260. Typed supplementation reaches 99.130, exceeds that ensemble candidate on every split, and earns acceptance in all nine (Table 3). Relative to the retained baseline, this removes 51.9% of the remaining pairwise ranking error,  $1 - \text{AUC}$ .

Table 3: **Typed inference adds value beyond a strong GNN/ensemble baseline.** Minesweeper, no label inputs; mean AUC (%) over nine official splits. Candidate mixtures are locked on selection data before independent acceptance. Rejected candidates return the retained baseline. The baseline menu and every paired result are in Appendix B.7.

| Path                          | Locked candidate | Deployed | Accepted |
|-------------------------------|------------------|----------|----------|
| Retained GNN/ensemble         | 98.189           | 98.189   | —        |
| Generic correction            | 97.530           | 98.189   | 0/9      |
| Independent-model ensemble    | 98.260           | 98.189   | 0/9      |
| <b>Typed inference (TSGL)</b> | 99.130           | 99.130   | 9/9      |

The typed specialist alone reaches 99.027 AUC. Its mixture with the baseline is better on average, illustrating the objective of supplementation: the system can retain useful information from both. The generic correction and independent ensemble are accepted on none of these splits. Their locked scores, not just their preserved deployment outputs, remain visible in the comparison. A descriptive six-model ensemble of both three-GNN fits does not recover the typed gain; it is not a ceiling on every possible ensembling procedure.

**Near-perfect prediction and an interpretable learned object.** In the ten-task integration panel, the matched strict result is  $97.552 \rightarrow 99.090$  AUC, and training-label anchors give  $98.432 \rightarrow 99.968$ . These remove 62.8% and 97.9% of the respective baselines’ remaining ranking error. This panel has two fits per access protocol, separate from the nine-split strong-baseline comparison.

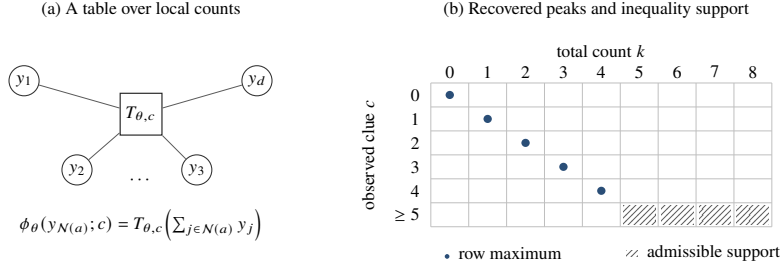


Figure 8: **Read the recovered law.** Equality-clue row maxima match the true count on exercised entries; the clipped clue recovers inequality support. The diagram reports maxima/support, not fitted probability magnitudes. Unexercised entries remain unidentified.

The cardinality-factor network (CFN) makes the learned relationship readable. Given the count family and clue membership, it fits entries in

$$\phi_{\theta}(y_{N(a)}; c) = T_{\theta,c} \left( \sum_{j \in N(a)} y_j \right).$$

It reaches .9906 AUC without label inputs and .9998 with training-label inputs (Table 4). Exercised equality rows peak at the true count; the clipped  $\geq 5$  row recovers inequality support (Figure 8). The supplied family determines what can be represented; the recovered entries show what was learned.

Table 4: **Minesweeper performance with declared access.** CFN uses ten official splits; strict TSGL here uses the nine-split strong-baseline protocol. Published GCN\* (Luo et al., 2024) and BuNN (Bamberger et al., 2025) provide external context, with their reported SDs. Cross-study mean differences are descriptive; the matched evidence is in Table 3.

| Model          | Information supplied                                    | AUC (0–1)         |
|----------------|---------------------------------------------------------|-------------------|
| Published GCN* | Training targets; no label inputs; generic architecture | .9786 $\pm$ .0024 |
| Published BuNN | Training targets; no label inputs; bundle diffusion     | .9899 $\pm$ .0016 |
| TSGL, strict   | Training targets; no label inputs; typed library/policy | .9913             |
| CFN            | Count family and membership; targets, no label inputs   | .9906             |
| CFN, anchored  | Same; training labels also supplied as inputs           | .9998             |

Minesweeper contains substantial deductive structure: sound propagation from the declared clues and training anchors logically entails 97.2% of test nodes. This is a property of the benchmark’s constraints and observation access, not certification of CFN predictions. With the numerical law supplied, exact inference in radius-three local models reaches about .9957 AUC without gradient training. Explicit relationships support both learning and execution.

## 6.2 PRESERVE CAPABLE PREDICTION ACROSS THE BENCHMARK PANEL

The ten-task integrated panel tests how selective supplementation combines with established perceivers. Across twenty fits and forty dependent evaluations, TSGL improves eight through structured inference and one through generic correction. In the remaining 31 it returns the saved baseline pointwise. All nine accepted system deployments improve on test (Table 5). On Roman-empire, retention preserves a strong predictor that already uses informative direction.

## 6.3 NATURAL DATA: VALUE BEYOND CORRECTION AND ENSEMBLING

**Preserve the evidence that inference needs.** Retaining direction under fixed feature-propagation recipes adds 6.57 accuracy points on Roman-empire and 9.85 on arxiv-year. These contrasts motivate strong directed baselines (Rossi et al., 2024). Appending the raw clue to the same Minesweeper representation adds 6.0 AUC points. The system therefore preserves raw observations alongside neural predictions before testing the value of further computation.

Table 5: **Validated gains or exact preservation in the ten-task panel.** Percent native metric; two fits and two access protocols per task. Upper group: nine gains and three Tolokers ties; lower: 28 exact ties. This panel uses two arxiv-year splits and two Minesweeper fits; the strong-baseline and ensemble comparisons have their own explicitly stated cohorts. Sources: [Platonov et al. \(2023\)](#); [Yang et al. \(2016\)](#); [Shchur et al. \(2018\)](#); [Lim et al. \(2021\)](#).

| Dataset                                                     | Metric | Strict: base → TSGL    | Anchored: base → TSGL  | Deployed path     |
|-------------------------------------------------------------|--------|------------------------|------------------------|-------------------|
| <i>Validated supplementation in at least one evaluation</i> |        |                        |                        |                   |
| Minesweeper                                                 | AUC    | 97.552 → <b>99.090</b> | 98.432 → <b>99.968</b> | Structured        |
| Arxiv-year                                                  | Acc.   | 63.440 → <b>65.235</b> | 64.409 → <b>66.473</b> | Structured        |
| Tolokers                                                    | AUC    | 84.308 → 84.308        | 85.117 → <b>85.409</b> | Generic (1/4)     |
| <i>Exact baseline preservation in every evaluation</i>      |        |                        |                        |                   |
| Roman-empire                                                | Acc.   | 92.932 → 92.932        | 93.223 → 93.223        | Baseline retained |
| Cora                                                        | Acc.   | 83.700 → 83.700        | 84.150 → 84.150        | Baseline retained |
| CiteSeer                                                    | Acc.   | 71.100 → 71.100        | 71.200 → 71.200        | Baseline retained |
| PubMed                                                      | Acc.   | 79.600 → 79.600        | 79.600 → 79.600        | Baseline retained |
| Amazon-Computers                                            | Acc.   | 91.945 → 91.945        | 91.873 → 91.873        | Baseline retained |
| Amazon-Photo                                                | Acc.   | 96.144 → 96.144        | 96.242 → 96.242        | Baseline retained |
| Questions                                                   | AUC    | 77.683 → 77.683        | 77.667 → 77.667        | Baseline retained |

**Arxiv-year.** The task uses the OGB citation graph ([Hu et al., 2020](#)) with LINKX year targets and official splits ([Lim et al., 2021](#)). Across four confirmation splits, TSGL beats both generic correction and a second-GNN ensemble (Table 6). Mean gains over the deployed ensemble control are 1.016 accuracy points strict and 1.559 anchored, positive on three and four splits, respectively; the ensemble wins the remaining strict split by .083 points. Against the locked *ungated* ensemble candidates (64.359/65.212), the differences are .980/1.363 points. The advantage survives both added-predictor and acceptance controls.

Table 6: **Arxiv-year: structure improves on correction and ensembling.** Mean accuracy (%) over four official splits of one graph. All branches share label access and the stated selection/acceptance protocol. Per-split scores, source-proposal reuse and budgets: [Appendix B](#).

| Access   | Baseline | Generic | Two-GNN ensemble | TSGL          |
|----------|----------|---------|------------------|---------------|
| Strict   | 63.556   | 63.700  | 64.322           | <b>65.338</b> |
| Anchored | 64.633   | 65.734  | 65.016           | <b>66.575</b> |

Useful supplementation need not mean standalone dominance. On two of these strict splits, the specialist scores 59.658, the baseline 63.672, and their mixture 65.442. The specialist checkpoint was selected for contribution to the mixture, which outperforms both components. This is the behavior TSGL is designed to exploit.

#### 6.4 A FROZEN RECIPE ON A NEW PATENT-CITATION GRAPH

We test transfer of the learning and deployment recipe on a previously unused patent-citation graph ([Lim et al., 2021](#)). The full graph exceeds the full-batch implementation’s memory budget, so the evaluation contract fixes three label-free induced views before fitting. Separately hash-seeded BFS unions produce 139,578, 134,299 and 157,171 vertices. Original features and directed edges inside each view are retained; year-bin targets use the full-source definition. Each view has one frozen train/selection/acceptance/test split.

The architecture, specialist library, membership policy, fitting recipe, mixture grid and acceptance rule are unchanged from the arxiv-year pipeline. Each model is normally fitted on the permitted patent training labels: this is transfer of a frozen recipe, not zero-shot parameter transfer. Generic correction and independent-model ensembling receive the same access.

Under strict access, typed mixtures beat generic correction by 1.47 points and the independent ensemble by 1.24, with both comparisons positive in every view. All three earn acceptance; deployment

Table 7: **Typed gains transport to a previously unused patent graph.** Test accuracy (%) on three predeclared induced views. Generic, ensemble and typed columns are locked mixture candidates. Every typed candidate is selected and accepted, so the typed column is also deployed TSGL. Views overlap and share one source graph; these are not official full-graph LINKX scores.

| View | Access   | Baseline | Generic | Ensemble | Typed / TSGL |
|------|----------|----------|---------|----------|--------------|
| 1    | Strict   | 66.27    | 66.37   | 66.77    | <b>68.15</b> |
| 1    | Anchored | 66.37    | 67.31   | 66.85    | <b>68.47</b> |
| 2    | Strict   | 64.76    | 65.27   | 65.50    | <b>66.37</b> |
| 2    | Anchored | 65.31    | 66.11   | 65.57    | <b>67.44</b> |
| 3    | Strict   | 65.06    | 65.58   | 65.64    | <b>67.12</b> |
| 3    | Anchored | 65.14    | 66.25   | 65.98    | <b>67.91</b> |

improves the baseline by 1.85 points on average. Anchored gains over the same controls are 1.38 and 1.81 points, with a 2.33-point improvement over the baseline. Both predeclared endpoints pass.

The result establishes useful transfer beyond the graph on which the recipe was developed. Its scope is equally specific: three overlapping, topology-selected views of one patent graph, within the citation-time task family. It is not a causal attribution of the gain to direction or an evaluation on a new mechanism family. Sampling, overlap, retained incidence and full masks are specified in [Appendix B.8](#).

#### 6.5 FRESH SOCIAL GRAPHS LOCATE THE TYPED ENGINE’S BOUNDARY

The same frozen pipeline is evaluated on three unused LINKX social datasets: Penn94, Twitch-gamers and Genius, each with three official splits and two access protocols. Across these 18 evaluations, the typed candidate beats the stronger of the generic and ensemble controls in **none**. Five typed proposals pass the acceptance test relative to the baseline, but the full system selects a generic or ensemble candidate instead. Thus passing a baseline-relative gate and providing the best available supplementation are different outcomes.

The full system records **14 improvements and four exact baseline returns**, with no observed regression ([Table 8](#)). The fourteen improvements comprise twelve generic-head and two ensemble deployments. These are useful system gains, not evidence of typed-inference superiority. The frozen evaluation reports the complete panel, including this absence of a typed win; it is not a task selection made after observing outcomes. The result bounds this library and recipe, not the presence of relational structure in the datasets.

Table 8: **System gains and typed gains are distinct.** Mean percent native metric across three official splits per access protocol. The full system improves fourteen cells and exactly preserves four. No typed candidate beats both locked controls. Per-cell candidates, accepted proposals and selected paths are in [Appendix B.8](#).

| Dataset       | Metric | Strict: base → system | Anchored: base → system | Gains | Typed wins |
|---------------|--------|-----------------------|-------------------------|-------|------------|
| Penn94        | Acc.   | 81.02 → 81.51         | 81.39 → 82.90           | 4/6   | 0/6        |
| Twitch-gamers | Acc.   | 65.11 → 65.39         | 65.83 → 66.29           | 4/6   | 0/6        |
| Genius        | AUC    | 90.96 → 91.72         | 90.96 → 91.74           | 6/6   | 0/6        |

**Cost and scope of selection.** Candidate fitting is a substantial cost even when deployment retains the baseline. In the ten-task panel, specialists account for about 4.88 of 5.42 recorded fitting hours, excluding development and other overhead. On Genius, the typed candidates consume 28.75 of 33.93 summed elapsed fitting hours on shared GPUs, yet are not selected in any of its six evaluations. These are per-fit elapsed times, not exclusive accelerator-hours or a matched hardware benchmark. The explicit baseline path avoids executing a refused specialist at deployment; it does not refund its training cost. Full accounting and failed development extensions are in [Appendix B](#).

## 7 LEARN A RELATIONSHIP ONCE; ANSWER MULTIPLE QUERIES

*In brief. Frozen learned count relationships support joint inference on independent boards without target fitting.*

TSGL tests whether relational computation improves a capable predictor. The compiler tests a second consequence: whether a learned relationship remains executable for subsequent questions. Node marginals alone do not determine joint-event probabilities. Direct node-to-joint exports and coherent prototypes did not yield useful learned dependence (Appendix C); the separate count-family compiler provides a positive test with independent target instances and explicit dependence controls.

Within a supplied count family, a source learner alternates an E-step over latent counts with an M-step for an observation table. We freeze the table and reuse it on 16 independent Minesweeper boards, of sizes  $100 \times 100$  and  $160 \times 160$ , at 50% and 25% anchor budgets. With grid membership and a local compiler supplied, the same object answers three event families: whether a pair contains two mines, whether a neighbourhood contains at least two, and whether it is safe. No target fitting or query-specific learning occurs.

Table 9: **Frozen relationships support new joint queries.** Macro Brier over three event families on independent boards; lower is better. Arms share accepted local support. Untrained-initialization and own-product controls separate learning from dependence.

| Predictor                                | 50% anchors | 25% anchors |
|------------------------------------------|-------------|-------------|
| Supplied numerical law                   | 0.019490    | 0.031530    |
| Learned, identity-initialized            | 0.019531    | 0.031670    |
| Learned, random-initialized              | 0.020949    | 0.033319    |
| Untrained identity initialization        | 0.029419    | 0.046795    |
| Learned model’s own product of marginals | 0.022196    | 0.038511    |

Both learned starts outperform the untrained initialization; the identity-initialized model closely tracks the supplied law. Joint predictions reduce Brier over products of its *own marginals* by 12.0% and 17.8%, with both board-bootstrap difference intervals above zero. This isolates dependence value while holding the individual marginals fixed.

Scores use common supported windows: 1,474 of 1,536 board/center/budget cells (95.96%), with abstention elsewhere. One sparse-anchor query family remains unresolved. From one source-fit regime, the compiler demonstrates local reuse at two sizes: one frozen object answers multiple questions on unseen boards.

## 8 RELATED WORK

The evidence–query–circuit decomposition connects several lines of graph learning. Our controlled experiments isolate the objects these methods change and test their alignment.

**Connectivity, reach and attention.** Graph diffusion changes propagation connectivity (Gasteiger et al., 2019b); over-squashing identifies bottlenecks in transmitting information (Alon & Yahav, 2021). Virtual nodes add communication paths (Gilmer et al., 2017), while GraphGPS separates local edge aggregation from global attention (Rampášek et al., 2022). These already make computation differ from raw adjacency. Our exact query pair asks: once observations are available, can a readout express the interaction needed for their optimal ranking? The answer differs between Bayes-matched queries on the same evidence. The result concerns fixed linear propagation and affine readout, rather than those architectures’ richer nonlinear operators.

**Learning structure and higher-order computation.** Neural relational inference learns latent interactions jointly with dynamics (Kipf et al., 2018); discrete structure learning optimizes an edge distribution for GCN prediction (Franceschi et al., 2019). Factor-graph neural networks represent multivariable dependencies and can express max-product belief propagation (Zhang et al., 2020). We

instead hold a declared circuit’s entire typed topology fixed while changing alignment, and cross membership with operators and observation access. This isolates connecting the right variables from choosing a more expressive operation. Learning those memberships extends the demonstrated separation.

**Learned inference and joint prediction.** Learned factor graphs fit local models for sum-product inference (Shlezinger et al., 2022); neural-enhanced BP learns message corrections (Garcia Satorras & Welling, 2021). GMNN couples neural representations with a conditional random field over labels, trained by variational EM (Qu et al., 2019). Our constructions test two consequences of this partnership: gains beyond access-matched correction and ensemble controls, and learned relationships reusable for event queries on independent instances. The compiler’s own-product control holds its node marginals fixed, isolating the value of retained dependence.

**Algorithmic alignment and reusable computation.** Algorithmic alignment studies how architectural structure matches a task’s computation (Xu et al., 2020); neural algorithmic reasoning develops neural execution of algorithmic processes (Veličković & Blundell, 2021). Our query intervention and fitted radius curves make one alignment requirement exactly testable. The compiler separates learning a numerical relationship from executing it on new instances within its supplied family.

**Strong prediction and selective correction.** Correct&Smooth exploits label correlations to post-process predictions (Huang et al., 2021). MLPs on fixed aggregation features rival trained GNNs on 12 of 14 benchmarks, with Minesweeper and Roman-empire as exceptions (Rubio-Madrigal & Burkholz, 2026); our interventions test which evidence and computation change their performance. TSGL evaluates structured gains against generic correction and independent-GNN mixtures. Its frozen recipe earns typed gains on patent-citation views but selects other branches on the social panel: a measured division of labor, not one inferred from separate modules alone. Further comparisons appear in Appendix H.

## 9 DISCUSSION

Observed relationships and computational dependencies are distinct structures. Our interventions establish the cost of identifying them. Bayes-matched queries on identical evidence require different interactions, and their fitted learning curves follow the exact ranking predictions. Recompile improves a common GCN update family; an isomorphic scramble that preserves the typed topology, architecture and initialization reduces performance to near chance. Alignment with the evidence, not graph shape alone, carries the gain. Correct observations can amplify a miscompiled circuit. Together these results establish circuit design as a substantive axis of graph learning, distinct from extending reach.

The constructive consequences are near-perfect Minesweeper prediction against a strong reproduced GNN/ensemble baseline, gains beyond generic correction and ensembling on both scholarly and patent-citation data, and learned relationships reusable across independent instances and queries. TSGL adds inference to strong predictors; the compiler retains executable relationships. The fresh social panel makes the distinction between system utility and engine superiority explicit: the system improves or preserves prediction, but its typed branch does not beat the alternatives. This is a boundary of the tested library, not evidence that those tasks contain no useful relationships.

Learning membership and building query-specific circuits automatically are the next architectural steps. The present experiments identify why those objects matter; they do not yet supply an unrestricted compiler or a unified node-and-joint learner. A general system should learn which relationships can be reused, execute them for the requested query, and retain a capable predictor when further inference does not earn its cost. The circuit and factorial interventions, strong-baseline comparison and fresh-data evaluations use preregistered endpoints and selection rules; frozen misses and post-hoc controls remain explicit in Appendix F.

The design principle is broader than either implementation: **treat the graph as evidence, and make inference an explicit object of design, learning and evaluation.** An observed edge tells us how the data are related. It does not, by itself, tell us how to answer the question.

## REPRODUCIBILITY STATEMENT

Proofs and their assumptions are in [Appendix A](#). Data access, graph construction, model selection, numerical scoring and all evaluation cohorts are specified in [Appendices B, C and E](#). The manuscript source contains the numerical records used for its tables and figures and independent checks of selected mathematical identities. Code is available on request. Prediction-retention exceptions, historical scorers and the distinction between numerical checks and trained-model reproduction are explicit in [Appendix H.2](#).

## REFERENCES

- Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory*, 2023.
- 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://openreview.net/forum?id=i800PhOCVH2>.
- Jacob Bamberger, Federico Barbero, Xiaowen Dong, and Michael M. Bronstein. Bundle neural networks for message diffusion on graphs. In *International Conference on Learning Representations*, 2025. URL <https://arxiv.org/abs/2405.15540>.
- Maya Bechler-Speicher, Ido Amos, Ran Gilad-Bachrach, and Amir Globerson. Graph neural networks use graphs when they shouldn’t. In *International Conference on Machine Learning*, 2024.
- Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M. Bronstein, Mathias Niepert, Bryan Perozzi, Mikhail Galkin, and Christopher Morris. Position: Graph learning will lose relevance due to poor benchmarks. In *International Conference on Machine Learning*, pp. 81067–81089, 2025.
- Zijun Cui, Hanjing Wang, Tian Gao, Kartik Talamadupula, and Qiang Ji. Theory-guided message passing neural network for probabilistic inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 667–675, 2024.
- Vijay Prakash Dwivedi, Ladislav Rampášek, Mikhail Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. Long range graph benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks*, 2022.
- Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *International Conference on Machine Learning*, 2019.
- Robert G. Gallager. Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1): 21–28, 1962.
- Víctor García Satorras and Max Welling. Neural enhanced belief propagation on factor graphs. In *International Conference on Artificial Intelligence and Statistics*, 2021.
- Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019a.
- Johannes Gasteiger, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning. In *Advances in Neural Information Processing Systems*, 2019b.
- 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>.

- William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.
- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. In *International Conference on Learning Representations*, 2021.
- Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, 2018.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
- Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Advances in Neural Information Processing Systems*, volume 34, 2021.
- Renming Liu, Semih Cantürk, Frederik Wenkel, Sarah McGuire, Xinyi Wang, Anna Little, Leslie O’Bray, Michael Perlmutter, Bastian Rieck, Matthew Hirn, Guy Wolf, and Ladislav Rampásek. Taxonomy of benchmarks in graph representation learning. In *Learning on Graphs Conference*, 2022.
- Yuankai Luo, Lei Shi, and Xiao-Ming Wu. Classic GNNs are strong baselines: Reassessing GNNs for node classification. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, 2024.
- Péter Mernyei and Cătălina Cangea. Wiki-CS: A Wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.
- Seth Nabarro, Mark van der Wilk, and Andrew J. Davison. Learning in deep factor graphs with Gaussian belief propagation. In *International Conference on Machine Learning*, 2024.
- Anthony Opipari, Chao Chen, Shoutian Wang, Jana Pavlasek, Karthik Desingh, and Odest Chadwicke Jenkins. Differentiable nonparametric belief propagation. *arXiv preprint arXiv:2101.05948*, 2021.
- Namyong Park, Ryan A. Rossi, Nesreen Ahmed, and Christos Faloutsos. MetaGL: Evaluation-free selection of graph learning models via meta-learning. In *International Conference on Learning Representations*, 2023.
- Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-GCN: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020.
- Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: are we really making progress? In *International Conference on Learning Representations*, 2023.
- Meng Qu, Yoshua Bengio, and Jian Tang. GMNN: Graph Markov neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5241–5250, 2019. URL <https://proceedings.mlr.press/v97/qu19a.html>.
- Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- Ladislav Rampásek, 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*, 2022.
- Tom Richardson and Rüdiger Urbanke. *Modern coding theory*. Cambridge University Press, 2008.

- Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M. Bronstein. Edge directionality improves learning on heterophilic graphs. In *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, pp. 25:1–25:27, 2024.
- Benedek Rozemberczki, Carl Allen, and Rik Sarkar. Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 2021. doi: 10.1093/comnet/cnab014. URL <https://academic.oup.com/comnet/article/9/2/cnab014/6271062>.
- Celia Rubio-Madrigal and Rebekka Burkholz. Fixed aggregation features can rival GNNs. In *International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=gSZhNPp103>.
- Oleksandr Shchur, Maximilian Mummé, Aleksandar Bojchevski, and Stephan Günnemann. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018. Relational Representation Learning Workshop, NeurIPS 2018.
- Nir Shlezinger, Nariman Farsad, Yonina C. Eldar, and Andrea J. Goldsmith. Learned factor graphs for inference from stationary time sequences. *IEEE Transactions on Signal Processing*, 70:366–380, 2022.
- Petar Veličković and Charles Blundell. Neural algorithmic reasoning. *Patterns*, 2(7):100273, 2021.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.
- Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. What can neural networks reason about? In *International Conference on Learning Representations*, 2020.
- Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 40–48, 2016.
- Zhen Zhang, Fan Wu, and Wee Sun Lee. Factor graph neural networks. In *Advances in Neural Information Processing Systems*, 2020.

## A FORMAL RESULTS AND THEIR EXACT SCOPE

### A.1 THE OBSERVATION MODEL AND AFFINE CONFINEMENT

Fix a Tanner graph with  $K$  observed check signs  $\sigma \in \{-1, +1\}^K$ . The latent bits are independent Bernoulli( $p$ ),  $0 < p < 1/2$ , and  $L_0 = \log((1-p)/p)$ . Check-node features contain their signs; bit-node features can contain fixed indicators. More generally, the entire feature array may be any affine function  $X(\sigma) = X_0 + \sum_{j=1}^K \sigma_j X_j$ . Propagation operators are fixed with respect to the observation realization. The readout is linear in any concatenation of propagated features, with an optional bias.

*Proof of Theorem 1.* For every fixed operator  $P$  and integer  $r \geq 0$ ,

$$P^r X(\sigma) = P^r X_0 + \sum_{j=1}^K \sigma_j P^r X_j.$$

Taking a node row, concatenating such rows and applying a linear readout preserves this form. The resulting score is  $S(\sigma) = \alpha + \beta^\top \sigma$ . The argument is unchanged for multiple fixed operators or any training procedure that chooses the readout coefficients. It also allows fixed affine standardization. No graph-size, radius or connectivity assumption is needed.  $\square$

The quantifiers matter. Once training has produced a score, the evaluation observation varies while the score's coefficients and preprocessing remain fixed. Normalization recomputed from that same observation can introduce an observation-dependent denominator and falls outside this argument. We do not claim that finite-sample empirical standardization differs by a proved  $o(1)$  term in the registered experiments. A theorem for that preprocessing requires additional assumptions and a separate concentration argument.

### A.2 WHEN AN AFFINE RANKING MUST LOSE AUC

Let  $\mu_1(a)$  and  $\mu_0(a)$  be the class-conditional probabilities of atom  $a$ , all strictly positive. For independent draws  $A_1 \sim \mu_1$  and  $A_0 \sim \mu_0$ , define

$$\text{AUC}(S) = \Pr\{S(A_1) > S(A_0)\} + \frac{1}{2} \Pr\{S(A_1) = S(A_0)\}.$$

A Bayes score ranks atoms by their likelihood ratio  $\lambda(a) = \mu_1(a)/\mu_0(a)$ .

**Lemma 1** (Extreme points of an affine score). *For  $S(\sigma) = \alpha + \sum_j \beta_j \sigma_j$ , maximizing coordinates satisfy  $\sigma_j = \text{sign}(\beta_j)$  whenever  $\beta_j \neq 0$ . Minimizers are their antipodes. In particular, a unique maximum has its unique minimum at the opposite corner.*

*Proof.* Each term  $\beta_j \sigma_j$  is maximized independently by this choice and minimized by its negative. A zero coefficient leaves that coordinate free, preventing a unique extremum.  $\square$

*Proof of Theorem 2.* For two atoms with  $\lambda(a) > \lambda(b)$ , ordering  $a$  above  $b$  rather than below it gains

$$\mu_1(a)\mu_0(b) - \mu_1(b)\mu_0(a) > 0$$

in their paired AUC contribution. Tying them loses half of that gain. Thus achieving Bayes AUC requires respecting every strict likelihood-ratio comparison.

In case (a), an affine score that does not put the unique highest atom  $u$  strictly first already loses. If it does, Lemma 1 places its unique minimum at  $-u$ . But the Bayes minimum is  $v \neq -u$ , so  $\lambda(-u) > \lambda(v)$  while the affine order reverses them. This also loses strictly.

In case (b), the distribution of  $S - \alpha$  is symmetric about zero under either class. Therefore  $S(A_1) - S(A_0)$  has a symmetric distribution, and its positive and negative probabilities agree. With half credit for ties, AUC is exactly  $1/2$ . A nonconstant likelihood ratio has a strict improving comparison, so Bayes exceeds  $1/2$ .

Finally, a finite atom set has finitely many weak orderings. If none of the affine orderings attains Bayes, their best AUC is separated from Bayes by a strictly positive gap. This establishes a gap for the supremum, not merely for each fixed choice of coefficients.  $\square$

Table 10: **Exact constructions and reported extensions.** The two higher-arity rows report numerical enumeration at the stated parameter, not a new closed form. At  $p = 1/20$ , the two-check Bayes AUC is 3971/4000, its gap is 171/4000, and the three-check Bayes AUC is 1598147/1600000. At  $\varepsilon = .05$ , the parity-gadget Bayes AUC is .905.

| Instance              | Bayes AUC                           | Affine comparison                |
|-----------------------|-------------------------------------|----------------------------------|
| Two-check chain       | $(1 - p)^2(1 + 2p)$                 | Cap $1 - p$ ; strict gap         |
| Three-check chain     | $(1 - p)^3(1 + 3p + 6p^2)$          | Strict gap by criterion (a)      |
| Arity (3, 3) chain    | Numerically enumerated              | Gap at least .0346 at $p = .05$  |
| Arity (4, 4) chain    | Numerically enumerated              | Gap at least .0280 at $p = .05$  |
| Direct two-check star | $(1 - p)^2(1 + 2p)$                 | Bayes attained; gap zero         |
| Noisy parity gadget   | $1 - 2\varepsilon(1 - \varepsilon)$ | Every affine score has AUC $1/2$ |

### A.3 AN EXPLICIT TWO-CHECK CALCULATION

Let the observations be  $\sigma_a = (-1)^{i+j}$  and  $\sigma_b = (-1)^{j+k}$  for independent bits  $i, j, k$ . In the leaf query, the target is  $i$ . The positive score orientation below ranks  $i = 0$  above  $i = 1$ ; negating the score gives the equivalent flip-probability orientation.

On this tree, the check-to-variable message from  $b$  to  $j$  is

$$2 \operatorname{artanh}(\sigma_b \tanh(L_0/2)) = \sigma_b L_0.$$

The next variable-to-check message is  $L_0 + \sigma_b L_0$ . Passing it through  $a$  multiplies it by  $\sigma_a$ , so the leaf posterior after two sum-product iterations is  $L_0(1 + \sigma_a + \sigma_a \sigma_b)$ . The product coefficient is nonzero, but the strict-gap result uses the ordering, as well as that algebraic observation.

For two signs, affine orderings are determined by the signs of  $\beta_a$ ,  $\beta_b$ ,  $\beta_a - \beta_b$  and  $\beta_a + \beta_b$ . The coefficient plane has eight open sectors, eight boundary rays and the origin: 17 weak-order classes. Enumerating the conditional atom masses gives the following seven possible values for  $(1 - p) - \operatorname{AUC}(S)$ .

| Slack below $1 - p$       | Number of weak-order classes |
|---------------------------|------------------------------|
| 0                         | 3                            |
| $1 - 2p$                  | 3                            |
| $(1 - 2p)/2$              | 3                            |
| $p(1 - p)(1 - 2p)$        | 2                            |
| $p(1 - p)(1 - 2p)/2$      | 2                            |
| $(1 - 2p)(p^2 - p + 1)$   | 2                            |
| $(1 - 2p)(p^2 - p + 2)/2$ | 2                            |

Every slack is nonnegative on  $0 < p < 1/2$ , and three orderings attain zero, including ranking by the target’s adjacent check sign alone. Hence the best affine AUC is exactly  $1 - p$ . Enumerating the Bayes order gives  $(1 - p)^2(1 + 2p)$ , proving Table 1. This finite calculation does not require a trained probe to find the optimal coefficients.

For the center query, the same two checks both touch the target directly. The posterior becomes  $L_0(1 + \sigma_a + \sigma_b)$  and is itself affine. Its AUC is the same Bayes value, with no affine gap. This control establishes the stated gadget boundary; it does not make all tasks of a specified graph depth equivalent.

The original verification protocol enumerates atom laws symbolically, compares AUC numerators over positive denominators, counts polynomial roots in the open parameter interval and checks signs at an interior point. It reports 15 of 15 checks passing. The retained compact source archive does not contain that original suite; `checks/audit_numbers.py` supplies separate independent enumeration and arithmetic checks (Appendix H).

#### A.4 A FINITE-CANDIDATE DISPATCH GUARANTEE FOR THE HISTORICAL RULE

This subsection analyzes the earlier empirical-argmax replacement rule. It is not a guarantee for the stricter paired-SE acceptance rule of the integrated node system in [Section 5](#). Exact fallback identity applies to both, but the near-best-candidate conclusion below requires the algorithm and IID assumptions stated here.

**Theorem 3** (Isolated dispatch under independent sampling). *Let  $h_0, \dots, h_K$  be predictors fixed independently of  $m$  identically distributed, independent dispatch examples. Let  $v(h)$  be their population accuracy and  $\hat{v}(h)$  the dispatch accuracy. Deploy an empirical maximizer when it strictly beats  $h_0$ ; otherwise deploy  $h_0$  exactly. For  $0 < \delta < 1$ , define*

$$\epsilon = \sqrt{\frac{\log(2(K+1)/\delta)}{2m}}.$$

*Then with probability at least  $1 - \delta$ ,  $v(h_{\text{deploy}}) \geq \max_k v(h_k) - 2\epsilon$ . In the refusal branch, predictions are pointwise identical to those of  $h_0$ .*

*Proof.* Condition on the candidate set. Hoeffding’s inequality and a union bound give  $|\hat{v}(h_k) - v(h_k)| \leq \epsilon$  simultaneously for all  $K + 1$  candidates. The rule selects an empirical maximizer even when it refuses: refusal means none strictly beats the baseline. Writing  $k^*$  for a population maximizer,

$$v(h_{\text{deploy}}) \geq \hat{v}(h_{\text{deploy}}) - \epsilon \geq \hat{v}(h_{k^*}) - \epsilon \geq v(h_{k^*}) - 2\epsilon.$$

The identity claim follows directly from returning the saved baseline artifact.  $\square$

The observed isolated study compares three saved seed checkpoints, so  $K = 3$ . An earlier calculation used  $K = 1$ ; correcting it enlarges the reported bound by a factor of approximately 1.08. For Reddit,  $m = 11916$ , the reported  $\epsilon$  is .0146 and the measured accuracy gap is .0032. These are empirical comparisons with the nominal bound. Application of the theorem to dependent graph-node outcomes requires justified independent units or a suitable dependence-aware inequality. Merely splitting node indices does not supply that assumption. AUC is pairwise and requires its own concentration analysis.

#### A.5 THE BALANCED NOISY-PARITY CONSTRUCTION

Let  $U, V$  be independent uniform signs, with target  $Y = UV$ . Observe  $\sigma_1 = UN_1$  and  $\sigma_2 = VN_2$ , where the independent noise signs satisfy  $\Pr(N_i = -1) = \epsilon$ ,  $0 < \epsilon < 1/2$ . Conditional on either target class, the two observation atoms of each parity are equiprobable, so the class-conditional laws are invariant under simultaneous sign reversal. The observed product matches  $Y$  with probability  $q = (1 - \epsilon)^2 + \epsilon^2$ . Its binary Bayes ranking has AUC  $q$  (including ties); every affine ranking has AUC  $1/2$  by [Theorem 2\(b\)](#). The gap is  $q - 1/2 = 2(\epsilon - 1/2)^2$ . This declared gadget is not a claim that arbitrary FSM observations have the required symmetry.

#### A.6 WHAT REMAINS CONJECTURAL

We conjecture that, in the measured law-governed families, low-access single-draw empirical risk minimization in generic message-passing classes tends to remain near the affine probe cap, whereas additional independent realizations can permit approach to a law-aligned reference. This is an optimization-and-access hypothesis, not a consequence of affine confinement. A general proof would need a specified model class, distribution, optimization procedure and asymptotic regime. Connections to parity and leap-complexity analyses are motivating analogies rather than transferred guarantees ([Abbe et al., 2023](#)).

#### A.7 FITTED READOUTS ON THE MATCHED QUERY CONSTRUCTION

For each of five draws (seeds 0–4), generate 20,000 independent triples of  $\text{Bernoulli}(p)$  bits and form a disjoint union of five-node paths  $i - a - j - b - k$ . Variable features are  $[1, 0, 0]$  and check features are  $[0, 1, \sigma]$ . Undirected edges connect adjacent positions in each path. With self-loops,  $P = D^{-1}(A + I)$  is fixed, and the readout bank concatenates  $X, PX, \dots, P^r X$ , for  $r = 0, \dots, 6$ . No empirical normalization is applied. Whole gadgets, rather than nodes within a gadget, are permuted into 50/25/25 training/validation/test sets. Both target fits receive the same feature array and split.

Linear and quadratic readouts are logistic regression; the latter appends every upper-triangular product of feature coordinates, including squares. The MLP has one hidden layer of width 32. Regularization is selected on validation AUC from  $10^{-4}, 10^{-3}, 10^{-2}$ . Logistic regression uses  $C = 1/(\lambda n_{\text{train}})$  and at most 5,000 solver iterations; the MLP uses at most 400 iterations and random state 0. The five draws vary the training observations, not the MLP initialization. The primary parameter is  $p = .2$ , with supporting values .05, .10, .30.

Each selected model is scored on held-out gadgets with tie-aware AUC and on the four sign atoms with exact population weights. For its fitted score  $f$ , the latter is

$$\sum_{a,b} \mu_1(a) \mu_0(b) \left( \mathbf{1}[f(a) > f(b)] + \frac{1}{2} \mathbf{1}[f(a) = f(b)] \right).$$

Atom weights enumerate the eight latent-bit configurations; they are not estimated from held-out frequencies. The graph feature bank is recomputed on the four sign atoms before evaluating the fitted readout. The signs are indistinguishable at radius zero; the leaf’s feature bank is independent of  $\sigma_b$  through radius two. Direct checks verify this property.

Balanced noisy parity uses independent uniform signs  $U, V$ , independent corruption signs with probability .10, and a three-node path with the target at its center. The two observed signs are on the leaves. A product readout can recover their parity from the squared sum once radius one reaches both leaves. This avoids supplying a nonlinear target encoding to the affine arm.

There are 945 selected readouts across query, radius, parameter, draw and head. Population optima and feature-reach identities are mathematical checks, separate from forecasts about finite-data fitting. Scalar fit records and timings are retained; fitted weights and per-atom score arrays were not saved. Repeating these inexpensive fits is therefore needed for an independent prediction-level replay.

## B INTEGRATED NODE-SYSTEM PROTOCOLS AND CONTROLS

### B.1 DEPLOYMENT ACCOUNTING ACROSS COHORTS

Each row in Table 11 compares the selected full system with its own saved, access-matched baseline. In every cohort, accepted system proposals improve the test metric and refusals preserve the baseline pointwise. A component can pass its own acceptance test without being the selected system proposal; those component-level acceptances are not counted as deployments.

Table 11: **Observed gains or exact preservation across all five deployment cohorts.** Counts refer to task/split/access evaluations. The ten-task panel already contains four arxiv-year evaluations reused in the ensemble comparison under its stated acceptance rule. Repeated splits, shared graphs and reused proposals preclude interpreting the rows as independent trials. The patent views belong to one source graph.

| Cohort                         | Evaluations | Improvements | Exact returns | Regressions |
|--------------------------------|-------------|--------------|---------------|-------------|
| Ten-task integration panel     | 40          | 9            | 31            | 0           |
| Strong-baseline Minesweeper    | 9           | 9            | 0             | 0           |
| Arxiv-year ensemble comparison | 8           | 8            | 0             | 0           |
| Patent-citation views          | 6           | 6            | 0             | 0           |
| Fresh social panel             | 18          | 14           | 4             | 0           |

The records cover fourteen source graphs, counting all patent views once and each reused benchmark once. We do not infer a population risk or universal no-harm guarantee from the absence of test regressions. Refusal has an exact algebraic guarantee; the performance of accepted proposals is an empirical result.

### B.2 DATA ACCESS AND FITTING

The ten-task integrated panel uses the released Minesweeper, Roman-empire, Questions and Tolokers masks (Platonov et al., 2023); public Planetoid masks for Cora, CiteSeer and PubMed; inherited

random masks for Amazon-Computers/Photo; and official LINKX arxiv-year splits. The Amazon masks allocate 60% training, 10% validation, 20% test and leave 10% unused. They must not be compared as if they were 60/20/20. Filtered squirrel in the stopped extension is the deduplicated benchmark, not the original unfiltered graph.

Each inherited validation mask is partitioned equally using seed  $99000 + s + 1000j$ , where  $s$  is the fit seed and  $j$  the split ID. Selection and acceptance are disjoint from each other and from training/test. All node features and edges are transductively available; only training labels are optimization targets. Selection labels choose checkpoints and mixture weights. Acceptance labels are read after proposals are saved; test labels score the deployment. These separations are per fit, not claims that historically studied datasets were unseen by the research program.

Table 12: **Actual confirmation fits.** IDs are split/seed. Counts are training/selection/acceptance/test nodes for the first fit in each dataset; masks are saved separately for every fit. Recipe columns show the first selected baseline, not the entire tuning menu.

| Dataset          | IDs      | Counts                  | Selected recipe             |
|------------------|----------|-------------------------|-----------------------------|
| Minesweeper      | 1/0, 2/0 | 5000/1250/1250/2500     | gcn, 12 layers, 64 hidden   |
| Arxiv-year       | 1/0, 2/0 | 84671/21167/21168/42337 | dir, 9 layers, 512 hidden   |
| Tolokers         | 0/0, 0/1 | 5879/1469/1470/2940     | gcn, 9 layers, 256 hidden   |
| Roman-empire     | 1/0, 2/0 | 11331/2832/2833/5666    | dir, 5 layers, 512 hidden   |
| Cora             | 0/0, 0/1 | 140/250/250/1000        | appnp, 1 layers, 512 hidden |
| CiteSeer         | 0/2, 0/3 | 120/250/250/1000        | appnp, 1 layers, 512 hidden |
| PubMed           | 0/0, 0/1 | 60/250/250/1000         | appnp, 1 layers, 512 hidden |
| Amazon-Computers | 1/0, 2/0 | 8251/687/688/2750       | sage, 4 layers, 64 hidden   |
| Amazon-Photo     | 0/0, 0/1 | 4590/382/383/1530       | gcn, 6 layers, 256 hidden   |
| Questions        | 1/0, 2/0 | 24460/6115/6115/12231   | sage, 6 layers, 512 hidden  |

The baseline menu reuses recorded GCN/GraphSAGE recipes with their learning rates, normalization, residual connections and feature preprocessing. Planetoid tasks additionally allow row-normalized features and APPNP (hidden 512, dropout .8, learning rate .01, weight decay .0005, budget 2000). Roman uses directed SAGE depth 5; official arxiv uses depth 9, width 512, batch normalization, residual connections, dropout .5, learning rate .001 and budget 2500. All winners are selected without acceptance labels. Released configuration records retain the complete per-fit menu and selected epoch; this is not a shared hyperparameter setting across every dataset.

Specialist inputs concatenate  $X, \hat{A}X, \hat{A}^2X$ . Directed graphs additionally provide incoming/outgoing one- and two-step features, including mixed compositions. This same raw/propagated feature bank enters generic heads and typed candidates. Generic heads have two hidden layers, width 128 or 256 and dropout .3. Typed candidates have width 128, three update families and ten rounds with damping .5. For  $C \leq 10$ , a factor takes at most eight distinct neighbours per role, selected by decreasing raw-feature cosine similarity after canonical edge deduplication; ties have deterministic ordering. A single affinity policy was selected over canonical membership on development CiteSeer, Minesweeper and arxiv-year cells. It is frozen for confirmation and invariant to edge storage order. For  $C > 10$ , four damped uncapped pairwise compatibility updates replace the higher-order engine. Thus the class-count branch changes the approximation, not just memory use.

Specialists use AdamW, learning rate .003, weight decay .0005, at most 600 updates, gradient clipping 1, selection checks every 10 steps and stopping after 200 steps without improvement once 200 steps have run. A decaying type-entropy term has initial coefficient .01 and vanishes at step 300. Every checkpoint is scored through its best mixture weight in  $\{.25, .5, .75, 1\}$ ; a reported standalone specialist is therefore *not* checkpoint-selected for standalone accuracy. In anchored fitting, half the training nodes are randomly exposed as observations per update and the rest are prediction targets. At evaluation all training labels may be inputs. The label-input perceiver is warm-started from the strict baseline and adapted for 300 steps. Correct&Smooth uses 50 correction and 50 smoothing steps and selection-chosen  $\alpha \in \{.5, .8, .95\}$ . The best of strict, label-input and C&S predictions becomes the preserved anchored baseline.

### B.3 LOCAL TABLES AND MESSAGE UPDATES

For a center label  $\ell$  and neighbour class channel  $c$ , the encoder emits local log-compatibility tables  $\theta_{v,\ell,c}(k)$ . The cardinality table indexes the count  $k$ ; the affine restriction is  $a_{v,\ell,c} + b_{v,\ell,c}k/D$ ; parity indexes  $k \bmod 2$ . These tables depend on the center’s raw/propagated features through the shared encoder. A learned task-level softmax mixes the three types in log-message space. It is a type mixture, not the perceiver–specialist mixture of Equation (3).

Cavity beliefs subtract the previous factor contribution. For a neighbour slot  $j$ , the count kernel computes the leave-one-out count distribution  $Q^{-j}(k)$  of the other binary class-channel indicators. After averaging exponentiated compatibility over the center’s cavity label belief, write the resulting table as  $G_c(k)$ . The message is

$$m_{c,j} = \log \sum_k Q^{-j}(k) G_c(k+1) - \log \sum_k Q^{-j}(k) G_c(k).$$

The parity kernel instead flips the parity index in the first contraction. Prefix/suffix contractions avoid enumerating all assignments at each update. The implementation also computes the center’s factor contribution, sums incoming messages into unary logits, and uses the declared damping and numerical clipping. Incoming and outgoing roles have separate table heads. Per-class binary-channel contractions and loops make the complete multiclass recurrence an approximation; exact local contractions do not establish a globally exact or coherent joint predictor. Anchors clamp permitted training-label logits during the recurrence. The alternative  $C > 10$  branch uses four iterations of a learned incoming/outgoing pairwise compatibility correction to its raw unary logits.

### B.4 SELECTION AND ACCEPTANCE

The system selects one best generic or structured proposal on selection data. It separately records the best generic control. For paired accuracy differences  $d_i$ ,  $\widehat{\text{SE}} = \text{sd}(d_i)/\sqrt{m}$ . For binary AUC, form each positive example’s ranking placement against negatives and each negative example’s placement against positives. For proposal-minus-baseline placement differences  $d_i^+, d_j^-$ , use

$$\widehat{\text{SE}}^2 = \frac{\text{var}(d^+)}{m_+} + \frac{\text{var}(d^-)}{m_-}.$$

Accept if the native-metric gain is at least .002 and greater than  $c\widehat{\text{SE}}$ , with  $c = 2.24$  for the ten-task panel. This formula does not account for arbitrary graph dependence; no coverage certification is asserted. Refusal returns the original saved array, verified pointwise. Replay of saved execution paths agrees to within  $9 \times 10^{-7}$  on the checked paths.

Table 13: **Component performance and contribution to the system.** Native metric in percent, two fits per dataset. Checkpoints were selected for mixture performance. These rows report each specialist separately from the deployed predictor; the complementarity comparison in the main text evaluates its contribution to the mixture.

| Dataset          | Metric | Strict specialist | Anchored specialist |
|------------------|--------|-------------------|---------------------|
| Minesweeper      | AUC    | 99.023            | 99.968              |
| Arxiv-year       | Acc.   | 60.928            | 65.040              |
| Tolokers         | AUC    | 81.696            | 80.831              |
| Roman-empire     | Acc.   | 46.294            | 73.641              |
| Cora             | Acc.   | 77.000            | 73.150              |
| CiteSeer         | Acc.   | 60.050            | 61.350              |
| PubMed           | Acc.   | 70.700            | 64.500              |
| Amazon-Computers | Acc.   | 90.873            | 76.345              |
| Amazon-Photo     | Acc.   | 78.562            | 84.706              |
| Questions        | AUC    | 68.147            | 65.417              |

### B.5 THE ARXIV ENSEMBLE COMPARISON

Reuse the saved baseline and specialist proposals on official confirmation splits 1–4. Train a second GNN from an independent initialization under the selected baseline recipe and the same masks; construct its anchored counterpart using the same label-input/C&S choices. Select its mixture weight on selection data. The controlled comparison uses  $c = 2.50$  for every reported branch, including the earlier saved generic/typed proposals. It does not change the original panel’s scores or frozen outcomes. Each new strict GNN has 9,533,445 parameters; additional fitting across its strict/anchored paths takes about 12–18 minutes per split. The comparison matches access, not parameters or total compute.

Table 14: **All paired arxiv confirmation results.** Accuracy in percent. Repeated splits and the two access protocols share one graph, so these eight rows are not independent graph populations.

| Split | Access   | Baseline | Generic | Ensemble | TSGL   |
|-------|----------|----------|---------|----------|--------|
| 1     | Strict   | 63.887   | 63.887  | 64.194   | 65.659 |
| 2     | Strict   | 62.992   | 63.247  | 64.894   | 64.811 |
| 3     | Strict   | 63.965   | 64.284  | 63.965   | 65.864 |
| 4     | Strict   | 63.380   | 63.380  | 64.235   | 65.019 |
| 1     | Anchored | 64.256   | 65.489  | 64.740   | 66.337 |
| 2     | Anchored | 64.563   | 65.784  | 65.612   | 66.608 |
| 3     | Anchored | 64.872   | 66.060  | 64.872   | 67.015 |
| 4     | Anchored | 64.839   | 65.605  | 64.839   | 66.341 |

Typed minus ensemble has split SD .8644 points strict and .4694 anchored. We report all paired values rather than infer a population interval from dependent split/access cells. The ungated comparison in the main text is a post-hoc diagnostic on already-locked proposals, with no refitting or reselection. Prior directed-engine results around 67.8% used different candidate fitting and selection; they are not substituted into this table.

### B.6 DIRECTION, REPRESENTATION AND STOPPED BREADTH

The earlier anchored direction-typed engine, separate from integrated TSGL, gives  $50.1523 \rightarrow 67.7842$  accuracy on five official arxiv-year splits when the factors separate citing and cited roles. Its paired mean gain is 17.6319 points with split-bootstrap interval  $[17.4627, 17.7873]$ . This official-split reproduction is descriptive; the seven-of-seven frozen prediction score belongs to the earlier random-split study. Both anchored arms fit training targets and use training-label observations. The official no-label-input direction-typed arm scores 65.0202; these means belong to this earlier fitting/selection protocol, not Table 6.

Fixed-perceiver direction contrasts use ten official Roman-empire splits, five official LINKX arxiv-year splits, and ten filtered squirrel/chameleon splits (Rozemberczki et al., 2021; Platonov et al., 2023). Roman gain is 6.57 points (paired-split bootstrap interval  $[6.28, 6.84]$ ), arxiv 9.85  $[9.69, 10.01]$ , squirrel 5.04  $[3.02, 7.19]$ , chameleon .50  $[-2.61, 3.34]$ . These intervals describe variation across splits of the same graph.

**Wikipedia record linkage and label-free verification.** The original recovery script matched filtered rows to geom-gen raw rows using exact binary feature vectors and dataset labels. All dataset labels entered these record keys, including validation/test nodes; they were not model inputs. A post-hoc reconstruction removes labels from both matching-key expressions and resolves duplicate feature vectors injectively by neighbourhood agreement with the filtered graph. It reproduces the saved raw-index mappings, directed edge arrays and split masks exactly for both datasets. Feature vectors alone uniquely identify 764 of 890 chameleon rows and 2,046 of 2,223 squirrel rows; the remaining candidates are resolved by observed graph structure. The resulting directed graphs have 13,516 and 65,577 edges, respectively.

Before repair, symmetrizing retained hyperlinks recovers all 46,998 squirrel edges and 8,853 of 8,854 chameleon edges; the single remaining chameleon edge is represented in both directions. The final undirected edge sets agree exactly. This check verifies connectivity consistency; equality to the saved

directed arrays additionally shows that labels were unnecessary to reconstruct the inputs actually evaluated. It is a post-hoc provenance check, not a new trained-model experiment or a revision of the frozen scores. The retained source archive includes the recovery implementation and its feature-only verification entry point.

On squirrel, a post-hoc same-recipe symmetrized Dir-SAGE control gives 36.6 versus 38.0 with direction,  $+1.37$  [.73, 1.99]. The tuned undirected incumbent versus a Roman-recipe directed model instead changes both recipe and direction and is not an isolating contrast. Chameleon has no resolved same-recipe direction effect. Direction-aware graph learning is established prior work (Rossi et al., 2024); these experiments measure its importance within this evidence/circuit decomposition.

The broad representation census was a component study, not the complete integrated system. A finite-budget replacement engine on tuned representations often loses. A separate exploratory fold-basis comparison evaluates the engine against the predictor that supplies its representation: on 15 of 18 tasks, the native-metric change is within one point. On Roman-empire’s directed fold basis, the earlier engine adds .12 points, with a paired-split interval [.01, .24]. These are measured increments from this engine, not bounds on all possible remaining headroom. The current directed baseline already retains the informative relation. The raw-clue bypass on Minesweeper gives 99.9 versus 93.9 AUC with the same representation retained. A linear reconstruction  $R^2$  is only a linear-readout statistic; these experiments do not prove that the representation contains no recoverable evidence. Benchmark design and protocol choices deserve separate attention (Liu et al., 2022; Bechler-Speicher et al., 2025).

The later filtered-squirrel and Amazon-ratings extensions used split 0 and seeds 10/11 for selection-only development. Continuation required typed mixtures to beat baseline, generic and ensemble controls in both seeds, with mean margin at least the larger of .003 and paired-seed spread. Neither task passed; acceptance and test values were not read and confirmation was not launched. No looser post-hoc rule rescues them. Recorded ten-task confirmation fitting is 19,507.9 seconds in total, with about 17,556.1 seconds on specialists and 1,951.8 on backbones; development, data preparation, repeated inference and other overhead are excluded.

**Published reference provenance.** The separate published tuned GCN reference is  $97.86 \pm .24$  AUC; Luo et al.’s final Table 3 reports GraphSAGE\*  $97.77 \pm .62$  (Luo et al., 2024). BuNN reports  $98.99 \pm .16$ , a ten-seed mean and standard deviation (Bamberger et al., 2025). Platonov et al. report GCN 89.75 and SAGE 93.51 and tune aggregation depth over 1–5 with their other stated hyperparameters fixed (Platonov et al., 2023). Published, project-rerun and access-matched integrated scores are distinct.

**Remaining-ranking-error reduction.** For the integrated Minesweeper comparison, the relative reduction is  $(\text{AUC}_{\text{TSG}} - \text{AUC}_0) / (1 - \text{AUC}_0)$ , using mean AUCs. Strict means .975515 and .9909025 give 62.8446%; anchored means .984315 and .999678 give 97.9471%. Ranking error here is  $1 - \text{AUC}$ , counting ties with half weight in positive/negative pair orderings. It is not node-classification error, and the ratio of mean errors is not a mean of per-split relative reductions.

## B.7 THE STRONG-BASELINE MINESWEEPER COMPARISON

Use official splits 1–9 with optimizer seed 0 for each primary GNN and seed 10000 for its independent counterpart. Split 0 is development-only. The fixed GNN configurations reproduce the reference GCN\*, SAGE\* and GAT\* implementations (Luo et al., 2024): hidden width 64, learning rate .01, weight decay 0, dropout .2, batch normalization and residuals, 2,000 epochs; depths are 12, 15 and 15, with one GAT head. Checkpoints maximize tie-aware AUC on the pre-partitioned selection mask. Baseline selection chooses among those three predictors and their equal-weight probability mean: the mean wins seven splits, SAGE one, GAT one. The independent-model control uses the same per-split menu choice formed from the independent fits.

Generic and typed candidate fitting, raw/propagated features, affinity membership, mixture grid and acceptance rule are unchanged. The typed library and generic heads are fitted on training labels only. Only the strict protocol is evaluated here; the backbone-specific anchored adaptation is not silently transferred to these reproduced models.

Table 15: **All strong-baseline paired results.** AUC in percent. Generic, ensemble and typed columns are locked candidates; every typed candidate earns acceptance. ENS3 is an equal-weight mean of the three GNN families. All nine splits share one board.

| Split | Baseline choice | Baseline | Generic | Ensemble | Typed  | Typed – ens. |
|-------|-----------------|----------|---------|----------|--------|--------------|
| 1     | ENS3            | 98.044   | 97.685  | 98.256   | 99.114 | 0.858        |
| 2     | ENS3            | 98.437   | 98.194  | 98.494   | 99.295 | 0.801        |
| 3     | SAGE            | 97.619   | 96.674  | 98.023   | 98.887 | 0.864        |
| 4     | ENS3            | 98.120   | 97.551  | 98.198   | 99.087 | 0.889        |
| 5     | ENS3            | 97.745   | 95.954  | 97.640   | 99.250 | 1.610        |
| 6     | ENS3            | 98.314   | 97.685  | 98.378   | 98.983 | 0.604        |
| 7     | ENS3            | 98.324   | 98.051  | 98.266   | 99.121 | 0.855        |
| 8     | ENS3            | 98.543   | 98.197  | 98.443   | 99.137 | 0.694        |
| 9     | GAT             | 98.555   | 97.778  | 98.646   | 99.294 | 0.648        |

The mean typed-minus-ensemble difference is .00869256, positive on all nine splits (range .006041–.016097). The deployed-minus-baseline mean is .00940844, with minimum .005945. Mean AUCs .98188844 and .99129689 imply a 51.9472% reduction in remaining ranking error. The typed specialist mean is .99027122; the mixture improves on it in seven splits, ties one and loses slightly on one. All specialist checkpoints were selected for mixture performance. A post-hoc comparison against an equal mean of all six GNN fits uses saved outputs and the frozen grid/rule; it remains descriptive, outside the frozen forecast score.

**Baseline scope and exposure chronology.** The baseline is the strongest reproduced member of this declared menu, not a matched realization of BuNN’s published .9899. The repository snapshot inspected for BuNN contained no implementation; a local paper-based reconstruction did not reproduce its reported level. That unsuccessful reconstruction makes no claim about the published result.

Baseline-preparation runners computed and logged held-out metrics. Selection itself used  $V_{\text{sel}}$  only; the recorded decision chronology identifies split 0 as the only held-out values exposed to the decision-maker, and excludes it from confirmation. The per-split baseline manifest was locked from selection scores before any TSGL fit. The accurate claim is selection-only model choice, not that held-out metrics were never computed. The source record preserves the logs and the preparation guard that suppresses these metrics.

## B.8 FROZEN-RECIPE NATURAL TRANSFER

The citation-time transfer and social panel share a frozen learning recipe: directed SAGE, width 512, depth 9, preprojection, residuals and BatchNorm, dropout .5, Adam at .001, zero weight decay, at most 2,500 epochs, minimum 500 and patience 300. Each inherited validation mask is halved with the same selection/acceptance rule. Both access protocols retain their own access-matched baseline, generic candidates, independently fitted GNN and typed candidate. All label values and all graph propagation used by a model belong to its declared observed graph.

**Patent views.** The source has 2,923,922 nodes. Construction indices  $v \in \{0, 1, 2\}$  correspond to the displayed views 1–3. For each  $v$ , hash the ASCII string formed by a fixed namespace, the view index and each decimal source node ID, with the exact string specified in the audit record in [Appendix F](#). Select the 17 smallest SHA-256 keys. On the full undirected support, with sorted neighbours, independently grow a FIFO BFS ball of at most 10,000 discovered nodes from each root. The view is their union, without refill. Sort retained nodes by source ID and retain every original directed edge with both endpoints inside. Each view has a 50/25/25 split made by NumPy permutation seed  $111000 + v$ , with validation then halved before fitting. View membership and masks were frozen before fitting or target summaries.

The 269 original features are retained. Class targets are the five full-source year quantile bins, not quantiles refitted within a crop. All propagations, factor memberships and anchors are recomputed

within the induced view. Thus no outside-view propagated feature or label observation is supplied to a candidate.

Table 16: **Patent-view construction and observation boundary.** Edges count unique directed pairs. Retained fractions are outgoing/incoming incidence fractions, not fractions of distinct edges touching a view. Counts are train/validation/test before validation halving.

| View | Nodes   | Edges   | Retained out/in | Split sizes          |
|------|---------|---------|-----------------|----------------------|
| 1    | 139,578 | 563,070 | 0.423/0.429     | 69,789/34,894/34,895 |
| 2    | 134,299 | 488,172 | 0.394/0.411     | 67,149/33,575/33,575 |
| 3    | 157,171 | 615,812 | 0.414/0.420     | 78,585/39,293/39,293 |

Pairwise node overlap is 17,735, 18,110 and 22,935. Views 1 and 2 each have one isolated root and one large weak component; view 3 has one weak component. These are overlapping, topology-selected observations of one new source graph, not independent graphs or an unbiased miniature of the full benchmark. Original stored duplicate edges are retained at intake and canonically coalesced by the propagation/membership operators.

The strict endpoint requires mean typed-candidate gain at least .002 over each control, superiority to both in at least two views, acceptance in at least two, and positive mean deployed-minus-baseline gain. Anchored access is scored separately by the same rule. Both meet all criteria; no sampling or recipe search follows an outcome.

**Social panel.** Penn94 and Twitch-gamers use accuracy; Genius uses ROC-AUC (Lim et al., 2021). Each uses official splits 0–2. Penn94 one-hot features omit its target metadata column; unlabeled nodes are excluded from every mask. Twitch targets the mature-account attribute and uses the reference standardized features; its mutual-follow graph is symmetrized because stored orientation is not a meaningful relation. Genius retains directed edges and its raw features. These choices were frozen before model fitting, together with the backbone recipe and resource contingencies.

Table 17: **All fresh social-graph results.** Percent native metric. Candidate columns precede acceptance; the final column names the actual deployed path.

| Dataset       | Split | Access   | Base   | Generic | Ensemble | Typed  | System | Path     |
|---------------|-------|----------|--------|---------|----------|--------|--------|----------|
| penn94        | 0     | Strict   | 80.989 | 82.174  | 82.020   | 81.566 | 80.989 | Base     |
| penn94        | 0     | Anchored | 80.989 | 82.875  | 81.999   | 81.597 | 82.875 | Generic  |
| penn94        | 1     | Strict   | 80.299 | 81.783  | 81.030   | 80.989 | 81.783 | Generic  |
| penn94        | 1     | Anchored | 81.422 | 82.947  | 82.442   | 82.246 | 82.947 | Generic  |
| penn94        | 2     | Strict   | 81.762 | 82.287  | 82.277   | 82.236 | 81.762 | Base     |
| penn94        | 2     | Anchored | 81.762 | 82.885  | 81.700   | 82.267 | 82.885 | Generic  |
| twitch-gamers | 0     | Strict   | 65.238 | 65.536  | 65.671   | 65.669 | 65.238 | Base     |
| twitch-gamers | 0     | Anchored | 65.774 | 66.318  | 65.726   | 66.242 | 66.318 | Generic  |
| twitch-gamers | 1     | Strict   | 64.372 | 64.796  | 65.214   | 65.041 | 65.214 | Ensemble |
| twitch-gamers | 1     | Anchored | 65.412 | 65.926  | 66.057   | 65.871 | 66.057 | Ensemble |
| twitch-gamers | 2     | Strict   | 65.712 | 65.840  | 66.164   | 65.809 | 65.712 | Base     |
| twitch-gamers | 2     | Anchored | 66.292 | 66.495  | 66.411   | 66.490 | 66.495 | Generic  |
| genius        | 0     | Strict   | 90.909 | 91.595  | 90.911   | 91.047 | 91.595 | Generic  |
| genius        | 0     | Anchored | 90.909 | 91.631  | 90.924   | 91.033 | 91.631 | Generic  |
| genius        | 1     | Strict   | 91.032 | 91.816  | 91.062   | 91.114 | 91.816 | Generic  |
| genius        | 1     | Anchored | 91.032 | 91.879  | 91.075   | 91.215 | 91.879 | Generic  |
| genius        | 2     | Strict   | 90.946 | 91.746  | 90.992   | 91.151 | 91.746 | Generic  |
| genius        | 2     | Anchored | 90.944 | 91.696  | 91.019   | 91.157 | 91.696 | Generic  |

Across eighteen cells, typed-minus-stronger-control is negative in every cell. Mean differences are  $-.006749$  on Penn94,  $-.001328$  on Twitch-gamers and  $-.006078$  AUC on Genius. Five typed proposals satisfy the baseline-relative acceptance test, but none wins full-system selection. Penn94

deploys four generic improvements and two exact baseline returns; Twitch deploys two generic and two ensemble improvements with two exact returns; Genius deploys six generic improvements.

The frozen positive rule requires superiority to both controls in at least four of six dataset/access cells and at least two typed acceptances. The frozen practically-null rule requires no typed acceptance and a small mean control contrast. All three datasets therefore have the literal classification *unresolved*. This classification does not obscure the observed absence of typed superiority over the controls.

## B.9 DEPENDENCE-SENSITIVE RESAMPLING

Use 2,000 bootstrap replicates of saved predictions, with one block-multiplicity draw shared across arms and overlapping splits. Recompute each split’s metric with those node weights, then average the paired contrasts. For AUC, tied scores are grouped before calculating weighted positive–negative comparisons. A numerical check agrees with standard weighted ROC-AUC. This is conditional sensitivity of fitted predictions; it does not include retraining or selection uncertainty or establish that graph blocks are independent.

Minesweeper uses  $5 \times 5$ ,  $10 \times 10$  and  $20 \times 20$  spatial tiles on the  $100 \times 100$  board. Across these scales, the 2.5/97.5 percentile intervals are, respectively:

| Contrast                         | Side 5        | Side 10       | Side 20       |
|----------------------------------|---------------|---------------|---------------|
| Aligned-eight – evidence-eight   | [.0380,.0480] | [.0384,.0472] | [.0377,.0481] |
| Aligned-eight – isomorphic       | [.4274,.4488] | [.4253,.4486] | [.4219,.4523] |
| Strong TSGL – ensemble candidate | [.0069,.0106] | [.0069,.0106] | [.0066,.0109] |
| Strong TSGL – baseline           | [.0075,.0114] | [.0074,.0115] | [.0073,.0117] |

These are resampling intervals on AUC differences, not graph-population confidence intervals.

Arxiv-year uses greedy BFS clusters on symmetrized support, starting in node-ID order, with target sizes 200 and 2,000. Strict typed-minus-ensemble candidate intervals are [.81,1.15] and [.80,1.15] accuracy points; anchored intervals are [1.18,1.53] and [1.17,1.58]. Realized partitions contain 31,619 and 29,428 clusters: about two thirds of nodes lie in large clusters, while roughly one quarter lie in fragments smaller than ten nodes. This fragmentation weakens interpretation of the nominal block sizes and is disclosed with the result. All declared primary contrasts remain above zero under their respective schemes; acceptance remains an empirical gate.

## B.10 RECORDED COSTS AND RETAINED OUTPUTS

The cost unit is the sum of per-fit elapsed times on shared, sometimes contended GPUs, not exclusive accelerator-hours or total study time. Recorded fitting hours are 5.7 for the circuit study; 2.2 for strong-baseline fitting plus .8 for its supplementation; 2.18, 6.91 and 33.93 for Penn94, Twitch-gamers and Genius; and 18.4 for the patent views. Genius’s typed candidates account for 28.75 hours. Gadget readouts record .1208 summed CPU elapsed-fit hours. The factorial retains scalar results without timings.

These totals omit some development, data preparation, interrupted-attempt waste and other overhead. Resource failures include out-of-memory attempts and a host-memory interruption; final fits resume from their cached completed stages without changing the recipe. Genius runs full-batch without the contemplated low-memory variant. All three patent views and all declared social cells complete.

Circuit and integrated-system probabilities and masks are retained, permitting independent metric replay. Gadget/FSM-readout fits and the replicated Minesweeper factorial retain scalar records but not full prediction arrays. On Genius, float32 rank accumulation in the original stored scorer differs from float64 standard ROC-AUC by at most  $6.8 \times 10^{-8}$ ; displayed conclusions and every acceptance decision are unchanged. The reported social tables here use standard tie-aware replay. Historical values are not silently overwritten.

## C INDEPENDENT-BOARD REUSE AND THE BOUNDARY ON JOINT INTEGRATION

### C.1 WHAT IS LEARNED AND FROZEN

Within the supplied clipped-count family, the observation model is a  $6 \times 6$  table  $T_{s,c} = \Pr(c \mid s)$  for latent clipped count  $s$  and observed clue  $c$ . Inference uses Bernoulli cell priors, anchor observations and local count chains, followed by exact variable elimination subject to a width cap. The source learner uses radius-two models for its E-step, six EM iterations and additive smoothing .1 in the M-step. The identity-biased initialization is  $.5I + .5/6$ ; the supplied-law comparator is  $I$ . Random initialization uses five Dirichlet row starts on 1000 source clue locations and selects by source evidence likelihood. The cell prior is estimated from permitted source training labels. These are source labels, not target-board fitting.

The independent-board study freezes one preselected source split’s learned tables and prior. It does not quantify source-fit uncertainty. Board seeds 981000–981015 generate 16 independent boards, eight at each size 100 or 160. Mine probability is .2, clue visibility .5, counts clip at 5, and nested independent uniform masks define the .5/.25 anchor budgets. Forty-eight centers per board are selected without labels, away from boundaries, with the pair-query members unanchored at both budgets. The target labels enter scoring only; the inference document receives labels only where the mask permits observations.

Each board/center/budget cell asks for a pair event, a neighbourhood count-at-least-two event and a zero-neighbour-mine event. Own-product controls use marginals from the same inferred distribution; the count control is a Poisson-binomial calculation. All arms try radius 3 and, on a width refusal, radius 2. If any arm refuses at that radius, all retry on the smaller common model. If neither succeeds, the cell is recorded as refused. There is no prediction or accuracy claim outside the common accepted support. Of 1536 cells, 1474 are accepted: 914 at radius 3 and 560 at radius 2. This denominator is not a count of independently sampled nodes.

### C.2 UNCERTAINTY AND CONTROLS

The independent unit is the generated board. Resampling boards retains all paired arms and events within each board. At 50% anchors, identity-initialized learned minus supplied-law macro Brier is .000041 with 95% interval [-.000055,.000133]. The random-initialized model is worse than the supplied law by .001459 [.000637,.002333]; useful learning does not make all initializations identical. Both learned starts outperform the *untrained identity-biased* comparator. This is not a comparison against each random start’s own untrained predictor.

Own-product minus learned-joint macro Brier is .002665 [.002185,.003133] at 50% anchors and .006840 [.005770,.007875] at 25%. The sparse-anchor pair family alone is unresolved. The size change in excess-over-supplied error is -.000002 [-.000207,.000192]. These are local reuse results, not proof of generalization to longer inference chains, unknown laws or arbitrary graphs. Recorded wall time is 114.97 seconds with 24 CPU workers; the reused source job’s 1694.31 seconds includes its earlier evaluation and is not a pure training cost.

An earlier study reused tables across masks on the same board. Source training masks included 1213–1274 of each target mask’s 2500 test labels; those cross-mask arms do not establish independent-instance transfer. Their same-split comparisons remain valid under declared access, but they are not pooled with this study. The independent generated boards here remove that source/target overlap.

### C.3 WHAT THE NODE-TO-JOINT ATTEMPTS DID NOT ESTABLISH

The first exporter interpreted node-trained parameters as a local energy and normalized it exactly. On Minesweeper, base-product Brier was .042116, the local model’s product .125674, and its joint .120661. The exported probabilities differed from the successful node predictor by about .318 in mean absolute probability difference, not Brier or accuracy. On Questions, local joint deployment was .085105 versus .076147 for the neural query control. The local interpretation did not reproduce the node model’s useful behavior; discriminatively trained potentials can define a joint distribution, but that fact alone does not establish a good one.

A later globally coherent forest mixture preserved its own marginals without earning useful structured deployment. The bounded Tolokers local model then used one stable sensor during fitting and execution and IPF to preserve its marginals. Of two seeds, one accepted a .25 mixture. Its query Brier was .125205, compared with .129181 for the full-perceiver product. A *post-hoc* no-refit control at the locked weight, removing learned interactions, scored .125107; the product of the deployed model’s own marginals scored .125071. The other seed refused and returned the baseline exactly. That deployment gain did not require learned dependence. Normalization and successful acceptance are not substitutes for this control.

The successful count-family compiler and these failed integrations have different training, context and inference procedures. They do not form a matched causal test of changing the objective alone. The general node/joint unification problem remains open.

## D PROTOCOLS, ACCESS DECLARATIONS AND ESTIMATOR HISTORY

### D.1 RESTRICTED DIAGNOSTIC AND ITS LIMITS

Write  $R(r)$  for the native-metric score of a fitted linear head on fixed propagated features at radius  $r$ . The historical rules use endpoint gain  $\Delta_R = R(r_{\text{last}}) - R(1)$  and reference headroom  $H_{\text{ref}} = R_{\text{ref}} - R(r_{\text{last}})$ . Where no computable reference exists and  $\Delta_R < .05$ , the rule labels saturation. Where  $H_{\text{ref}} \geq .05$  and  $\Delta_R < .02$ , it labels plateau; otherwise slope. These native-metric thresholds are not universal constants. A low radial gain does not establish closeness to the best possible learner, and an unavailable reference is undefined, not zero.

A registered four-configuration GNN menu supplies a finite comparator, not a ceiling. Circuit-cone and epidemic-time statistics are query-matched diagnostics, but not necessarily affine functions of the same sign encoding. Only the specified linear members satisfy [Theorem 1](#). Correct&Smooth stays within the registered envelope plus spread on eight of nine tested tasks; Actor is the exception. Fixed GDC improves the tested plateau probe by only.002. Neither observation is an impossibility result for learned rewiring or nonlinear graph models.

Effective realizations (ER) count declared local-law supervision: labeled bits times independent draws for LDPC, or observed adjacent anchor pairs for the FSM. They are family-specific counts, not mutual information. A measured FSM sweep orders together under matched counts, while a proposed universal per-table-entry normalization fails ([Appendix E](#)). Classical decoding and trellis references follow their stated observation models ([Gallager, 1962](#); [Richardson & Urbanke, 2008](#); [Rabiner, 1989](#)). Where exactness is not established, they remain principled references rather than Bayes optima.

### D.2 NAME THE PROBE THAT PRODUCED THE RESULT

| Version        | Optimization / metric                | Role in the record                                                       |
|----------------|--------------------------------------|--------------------------------------------------------------------------|
| RProbe-v1      | LBFGS, at most 200 steps; binary AUC | Original 20-cell functional-validation panel                             |
| RProbe-v2      | Adam, 300 steps; binary AUC          | Prospective panel on 18 tasks                                            |
| LRGB extension | Adam, 300 steps; macro-F1            | PascalVOC-SP ( <a href="#">Dwivedi et al., 2022</a> ); radii 0 through 8 |
| RProbe-public  | Adam, 300 steps; native metric       | Named package interface with provenance object                           |

The ordinary radius set is  $\{0, 1, 2, 3, 5\}$ ; factor-circuit studies additionally reach radius 99. The node and factor probes use symmetric normalization of  $A + I$ . In v2, weight decay is validation-selected from  $\{10^{-4}, 10^{-2}\}$ . Circuit cells use a directed-forward-cone statistic; epidemic cells use radius-restricted infection-time statistics. Those estimators are related diagnostics but are not all members of the affine class of [Theorem 1](#).

Finite fitted curves can decrease with radius. The alternative estimator  $\hat{R}^*(r) = \max_{s \leq r} R(s)$  changes exactly one of eighteen frozen regime labels: squirrel, whose original  $\Delta_R = .0499$ , moves to slope. We retain the original scoring rather than retrospectively treating the monotone variant as the preregistered instrument.

Legacy LDPC weight-decay selection used the evaluation split. A correction uses a 40/20/40 train/validation/test split and validation-only selection. The  $(4, 8)$ ,  $n = 1200$  cell gives  $\Delta_R = .0004$  and reference headroom .086; the  $(3, 6)$ ,  $n = 4800$  cell gives  $\Delta_R = -.0013$  and headroom .059. Both retain the plateau label. The corrected labels do not make the original selection procedure valid.

### D.3 FUNCTIONAL VALIDATION: RETAIN THE PROVENANCE LIMIT

The earlier instrument-validation gate concerns the area *above* the curve, not its endpoint gain. On a 20-cell panel of 12 circuit cells and 8 datasets, the recorded rank correlation against internal depth-structure values is .908 for the area functional and .086 for the endpoint. The original executable scorer did not survive.

The deterministic replacement audit uses a mechanically sourced target vector from registered run files. Nineteen cells are usable; one dataset lacks a registered target. Its Spearman correlations are .82 for area and .26 for endpoint. The leave-one-circuit-out range is .80–.82; circuits alone give .62; datasets alone give  $-.71$  on targets with very little variance. This supports mixed-panel ranking, not a common within-family effect. Neither the area functional nor the original .908 defines the frozen regime labels; those use radial gain and reference headroom.

### D.4 THE GENERIC ENVELOPE AND INTERACTION PROBES

The registered generic menu contains four configurations: GCN with three layers; GraphSAGE with three layers; a family-specific deeper/wider GraphSAGE (four layers and hidden width 512, or nine layers); and APPNP (Kipf & Welling, 2017; Hamilton et al., 2017; Gasteiger et al., 2019a). Each is screened for 300 steps. The validation winner runs three seeds at the full budget with best-validation checkpointing and patience. Envelope spread denotes the reported maximum-minus-minimum at convergence; it is not a confidence interval or a universal architecture envelope.

The nonlinear probe is a 64-unit ReLU hidden layer on the same  $F_r$  as the linear probe. It uses Adam at learning rate .01 for 300 steps, validation-selected weight decay from  $\{10^{-4}, 10^{-2}\}$ , per-radius standardization and seed 0. The quadratic probe uses a fixed seed-0 random projection to 64 dimensions, appends all 2080 pairwise products and fits a logistic head under the same optimization protocol. Neither trains a graph representation. Post-hoc replications use seeds 0–9; seed 0 reproduces three of four examined frozen task values exactly, and the frozen WikiCS value lies inside the replicated seed range.

### D.5 TRANSFORMER CONFIGURATIONS AND THE FITTING CHECK

The registered GPS-class arm (Rampásek et al., 2022) uses features plus eight Laplacian positional coordinates; three GPSConv layers of hidden width 64; a GraphSAGE local block; two-head global attention; dropout .3; and a linear head. Optimization is Adam at .001, weight decay  $10^{-5}$ , validation checkpointing and patience, with three seeds. Four further variants change one architectural choice: six layers; width 128 with four heads; sixteen positional coordinates; or removal of the local block (attention-only).

Thus the detailed protocol specifies *five total configurations*, including attention-only. The stronger fitting diagnostic doubles capacity and tests learning rates .003 and .01 for 3000 steps. Both variants reach training-anchor accuracy 1.0000, but test accuracies fall to .0978 and .0502. This rules out an explanation based solely on inability to fit those anchors; it does not prove that every optimization schedule must fail.

### D.6 TASK FAMILIES AND OBSERVATIONS

**LDPC decoding** Regular  $(d_v, d_c)$  Gallager ensembles use a configuration model without duplicate edges. Independent bit flips have probability  $p = .05$ . The observations are syndromes and the target is bit-flip ranking, scored by AUC. Syndrome-domain sum-product runs 50 iterations. It is exact on the tree gadgets and a principled reference on loopy ensemble graphs. The access count is labeled training bits multiplied by independent draws.

**Finite-state machines** A random deterministic transition table  $\delta$  has  $C$  states and  $S$  symbols. A length- $n$  walk emits observed symbols corrupted with probability  $\varepsilon = .05$ . Anchors reveal states at density  $\rho$ . The ordinary target is non-anchor state recovery. A trellis forward-backward solver supplied with the true  $\delta$  and  $\varepsilon$  provides the exact reference. The clean query intervention additionally reveals true symbols at the same anchors to both query arms. Its access count is observed adjacent anchor pairs.

**Circuit faults** ISCAS’85 netlists are subjected to single stuck-at faults, with observation density .05. Signature matching identifies a fault exactly only when its observed signature is unique. Both selective accuracy and all-detected-fault metrics must be reported with their denominators.

**Epidemic sources** The observation process is discrete susceptible-infected (SI), not SIR, with sparse observation times on real topologies. The target is source finding, scored by MRR. Dynamic message passing is a principled reference whose performance can itself be limited by the observation process.

**Node benchmarks** The development suite spans 15 standard benchmarks, including Cora, WikiCS, Actor, WebKB tasks and Reddit (approximately 233,000 nodes and 41 classes), loaded through PyTorch Geometric (Mernyei & Cangea, 2020; Pei et al., 2020; Hamilton et al., 2017; Fey & Lenssen, 2019). The prospective containment subset has eleven node tasks. These counts describe different panels, not competing totals for one panel.

#### D.7 WHAT EACH ARM IS TOLD

| Label | Supplied information                                                      | Typical arm                          |
|-------|---------------------------------------------------------------------------|--------------------------------------|
| L0    | No law family or law parameters                                           | Generic end-to-end network           |
| L1    | Typed library or declared factor family; infer permitted types/parameters | Learned typed engine / CFN           |
| L2    | Factor scaffold; infer model parameters                                   | Compiled-circuit GCN                 |
| L3    | Governing law and parameters                                              | Trellis, BP, DMP, signature matching |

These labels expose information differences; they do not turn the comparison into an equal-information architecture contest. They are access declarations rather than a formally nested hierarchy. In particular, Minesweeper’s supervised no-label-input CFN receives the cardinality family. Supplying a family and learning its parameters is different from discovering an arbitrary factor family without prior structure.

### E MEASUREMENT PROTOCOLS AND CONTROLLED COMPARISONS

#### E.1 CIRCUIT COMPILATION AND ALIGNMENT

Use PyTorch Geometric’s Minesweeper dataset (Platonov et al., 2023), with 10,000 variable nodes, seven one-hot feature channels and the official 50/25/25 masks for splits 0–4. The symmetrized simple graph has 39,402 undirected edges. Feature zero denotes an unrevealed clue; the other six channels encode counts 0, 1, 2, 3, 4,  $\geq 5$ . Exactly 5,000 clues are revealed.

**Observed data and supplied semantics.** For each revealed host  $v$ , create a factor  $f_v$  connected to  $N_E(v)$ , excluding  $v$  itself. The compiled graph has only variable-factor edges: 15,000 nodes and 39,404 undirected incidences. Variables retain all seven features; factors receive their host’s six observed count channels. No hidden count or target label is consulted. The scaffold supplies the interpretation that a clue constrains its neighbours. It is derived deterministically from the evidence under that interpretation; the experiment does not discover those semantics.

**Architecture and training.** All arms propagate using  $P = D^{-1}(A + I)$  and residual updates

$$h^{\ell+1} = h^{\ell} + \text{Drop}_{.2}(\text{ReLU}(\text{LN}(P \text{Linear}_{\ell}(h^{\ell})))) .$$

The evidence arm applies one input linear map and ReLU. The bipartite arm has one map per node type, a learned type embedding and ReLU, followed by the shared update family and a two-logit head. Factor nodes are excluded from fitting and evaluation. Widths keep all parameter counts within .7% of 135,426.

Fit with full-batch cross-entropy, AdamW at learning rate .003, weight decay .0001, at most 1,500 updates, and validation checks every ten updates. Stop after more than 300 updates without validation improvement and restore the selected state. Both selection and test use standard tie-aware ROC-AUC. The primary evidence/aligned/scrambled comparisons use eight layers, five splits and optimizer seeds 0–4; the sixteen-layer aligned arm is a separate higher-depth reference.

Table 18: **Depth and parameter accounting for the circuit comparison.** Means and descriptive SDs across the stated fits. Split/seed evaluations share one board and do not constitute independent graph populations.

| Circuit                 | Layers | Width | Parameters | Fits | AUC    | SD     |
|-------------------------|--------|-------|------------|------|--------|--------|
| Evidence adjacency      | 8      | 128   | 135426     | 25   | 0.9021 | 0.0033 |
| Aligned factors         | 8      | 127   | 134495     | 25   | 0.9449 | 0.0033 |
| Isomorphic scramble     | 8      | 127   | 134495     | 25   | 0.5075 | 0.0126 |
| Degree-preserving swaps | 8      | 127   | 134495     | 10   | 0.5058 | 0.0104 |
| Aligned factors         | 16     | 90    | 135632     | 25   | 0.9550 | 0.0039 |

**Randomization and invariance.** For the isomorphic control, permute variable endpoints within equal-incidence-degree classes using seed  $107000 + k$ . Hold factor features and variable features, labels and masks fixed. The primary combined replicate pairs wiring seed  $k$  with optimizer seed  $k$ , and with the corresponding aligned optimizer seed. Five extra wiring seeds 5–9 at split 0 and optimizer seed 0 measure wiring-only variation. Membership overlap is .09–.24% and wiring-only AUC SD is about .013. A separate eval-mode co-permutation of graph and variable features changes predictions by at most  $9.9 \times 10^{-7}$ .

The degree-swap control exchanges endpoints between factor incidences, rejecting duplicate edges and preserving both degree sequences. It uses two seeds on each of five splits. This randomization changes topology, unlike the isomorphic permutation.

The mean aligned-minus-evidence gain is .04277848, positive in all 25 pairs; split-averaged gains range .03805–.04694. Aligned-minus-scrambled is .43740978, also positive in every pair; split means range .42758–.44691. These ranges describe the evaluated board. Spatial-resampling sensitivity is reported with the evaluation protocols in [Appendix B.9](#).

**Historical controls and score provenance.** The archived single-seed study reports evidence .903524, compiled-eight .943693, compiled-sixteen .955917, random-membership-sixteen .505651, clique completion .742937, and evidence-edge rewiring .508880, over five splits. Its random membership preserves factor degrees but not variable degrees. Clique completion adds pairs within revealed neighbourhoods; rewiring preserves variable degrees. The historical rank scorer did not average ties, and original prediction arrays/checkpoints were not retained. Those values cannot be rescored from scalars. The primary numbers above are new tie-aware fits, not a claimed correction of the archived outputs.

**Cora control and deduction strata.** In the earlier Cora control, replacement changes accuracy by  $-3.3$  percentage points and augmentation by  $+.25$ . Its code uses seed-defined 60/20/20 masks, factor features  $\log(1 + \deg v)$ , and keeps the original edges only in the augmentation arm. Absolute Cora scores and uncertainty are not substituted from the later integrated panel.

Nodes can also be stratified by the round at which a specified sound constraint-propagation procedure determines them. Reported GCN accuracy falls approximately  $.89 \rightarrow .78 \rightarrow .70$  across rounds 1–3, with similar accuracy across the tested 2-, 4-, 8- and 16-layer models. The available summary does not supply four independent curves or justify statistical equivalence, so we do not draw them. With anchors clamped, the aligned engine remains at 1.0 through round 4+. These rounds are not proved minimum depths of all algorithms; anchor availability, clue density and other stratum differences

can also vary. They distinguish the stated procedure’s composition requirements from receptive-field reach without identifying every cause of the stratified gap.

## E.2 MEMBERSHIP, OPERATOR AND ANCHOR FACTORIALS

The Minesweeper factorial uses five official splits and optimizer seeds 0–4. It holds the randomized membership realization at seed 0, so these are optimizer/split replications, not 25 independent random circuits. The FSM factorial uses five draws and a fixed solver. All sixteen means are shown below.

Table 19: **Complete factorial means.** Minesweeper: 25 split/optimizer fits per cell, tie-aware AUC. FSM: five traces, fixed belief-propagation solver, accuracy. Each panel retains its own observation and membership construction.

| Membership | Operator | Minesweeper AUC |        | FSM accuracy |        |
|------------|----------|-----------------|--------|--------------|--------|
|            |          | Clamped         | None   | Clamped      | None   |
| True       | Typed    | 0.9997          | 0.9909 | 0.7792       | 0.7414 |
| True       | Affine   | 0.8906          | 0.8872 | 0.1443       | 0.1233 |
| Random     | Typed    | 0.5043          | 0.6964 | 0.1473       | 0.1483 |
| Random     | Affine   | 0.5163          | 0.5301 | 0.1424       | 0.1233 |

**Minesweeper.** The count-factor model has width 128, neighbourhood cap 8, ten inference rounds and damping .5. Random membership permutes rows of the neighbour-list matrix while leaving host features fixed. Typed tables are unrestricted over center label and neighbour count; the affine arm restricts count dependence to an affine function. Use AdamW at .003, weight decay  $10^{-5}$ , binary cross-entropy, gradient clipping 1, at most 600 updates, validation every ten updates, and patience greater than 200. Both validation and test AUC are tie-aware; the test value attached to the validation-best checkpoint is retained. Clamped fitting reveals a random half of training labels and predicts the remainder, with clamps  $\pm 12$ . Validation/test execution reveals all training labels. The no-anchor arm fits all training targets without label inputs.

The clamped typed-minus-affine gain is .1091 under true membership and  $-.0120$  under random membership. Clamping changes the typed arm by  $+.0088$  under true membership and  $-.1921$  under random membership. Both signs hold in all 25 pairs. These models retain scalar rows, not prediction arrays or checkpoint tensors. The archived seed-0 anchor drop of  $-.2041$  and its frozen prediction remain part of the forecast record; they are not substituted for the 25-fit means.

**Finite-state machine.** Use an  $(8, 4)$  transition table generated with seed 4242 and traces with seeds 0–4, length 20,000, anchor probability .02 and symbol corruption .05. True factors connect  $(t, t + 1)$ ; random factors connect  $(t, \rho(t))$  for a seed-0 permutation, dropping self-pairs. This randomization does not preserve chain topology. Typed potentials use the supplied transition table and channel likelihood. Affine potentials are  $\theta \mathbf{1}[s_u = s_v]$ , with  $\theta \in \{.25, .5, 1, 2\}$  selected on validation. All arms use uniform messages, 60 log-domain BP iterations and damping .5. Anchors clamp state unaries where permitted. No gradient fitting occurs.

With anchors, typed-minus-affine is .6349 under true membership and .0049 under random membership. Clamping changes the typed arm by  $+.0378$  and  $-.0010$ , respectively. The true/typed/clamped accuracy .7792 agrees closely with an independently implemented trellis value .7789 in the wrong-law study; that cross-implementation agreement does not prove every finite-iteration BP marginal exact.

## E.3 STATE AND SYMBOL QUERIES ON ONE EVIDENCE ARRAY

Generate an FSM with uniformly drawn transition-table entries, uniform initial state and uniform symbols. Evolve  $s_{t+1} = \delta(s_t, x_t)$  for 20,000 positions. Each symbol is independently corrupted with probability .05 by replacement with a different symbol. Select observation anchors independently with probability .02. Four draws have  $(C, S) = (8, 4)$  and seeds 4242, 9101, 9102 and 9103; the fifth has  $(16, 4)$  and seed 9201. The probe graph is the undirected time path, with self-loops added by normalization.

**Identical evidence; disjoint roles for labels.** Let  $a_t$  indicate membership in the observation set  $O$ . Both questions receive the single tensor

$$X_t = [\text{onehot}(\tilde{x}_t); a_t \text{onehot}(s_t); a_t \text{onehot}(x_t); a_t].$$

Its width is  $2S + C + 1$ . Outside  $O$ , both true-value channels are zero. Split non-anchor positions into a provisional validation half and test half. Draw  $|O|$  supervised positions uniformly from the former, using seed 106000+draw seed, to form  $S$ ; its remainder is  $V$ . Train on  $S$ , select on  $V$ , and score the untouched test half  $T$ . The observed anchor set and the supervision set never overlap. The same  $O, S, V, T$  apply to both questions. Access is explicitly  $|O|$  observed state/symbol pairs plus  $|S|$  target labels.

**Fitting and reference.** Use  $P = D^{-1/2}(A + I)D^{-1/2}$  and concatenate  $X, PX, \dots, P^r X$  for  $r \in \{0, 1, 2, 3, 5\}$ . Column means and scales are computed on  $S$  only, with a  $10^{-6}$  scale floor. Fit each linear softmax head for 300 Adam steps, learning rate .01, seed 0, with weight decay  $10^{-4}$  or  $10^{-2}$ . Choose radius and weight decay by validation accuracy, not test accuracy. A three-layer width-256 residual GCN uses BatchNorm, dropout .5, Adam .01, gradient clipping 1, up to 800 updates, validation checks every ten updates, patience over 200, and seeds 0/1.

The forward-backward reference receives the true transition table, corruption law, uniform priors and the same observation anchors. Revealed states and symbols clamp the respective messages. It does not receive labels at  $S$  as extra observations. Both posterior targets are read from this common reference calculation. The symbol question is near-observed; it is not Bayes-difficulty-matched to the state question.

Table 20: **Identical evidence and disjoint observation/supervision masks.**  $|S| = |O|$ ;  $r_s$  is the validation-selected state-probe radius. The means are fitted empirical accuracies, not a theorem about finite-data probe fitting.

| Draw | $(C, S)$ | $ O $ | $r_s$ | State probe | Reference | Symbol probe | Reference |
|------|----------|-------|-------|-------------|-----------|--------------|-----------|
| 1    | (8,4)    | 397   | 5     | 0.2726      | 0.7827    | 0.9512       | 0.9515    |
| 2    | (8,4)    | 397   | 2     | 0.2251      | 0.7048    | 0.9489       | 0.9494    |
| 3    | (8,4)    | 376   | 2     | 0.3217      | 0.8173    | 0.9512       | 0.9515    |
| 4    | (8,4)    | 393   | 2     | 0.2492      | 0.7375    | 0.9504       | 0.9514    |
| 5    | (16,4)   | 411   | 2     | 0.1758      | 0.6541    | 0.9490       | 0.9494    |

The mean state deficit is .4904, with range .478–.510; the mean symbol deficit is about .0005. Scalar probe curves and selected radii are retained, but fitted weights and prediction arrays are not. The corresponding historical shared-array protocol trained at observed-anchor rows and reported the test-best radius. Its numerical record and frozen score are not rescored here; it is not the protocol underlying Figure 6. Its shortcut contribution cannot be isolated by subtracting results across protocols that also change supervision and selection.

#### E.4 INTERACTION-ORDER PROBES AND POST-HOC SEED STUDIES

Table 21: **Changing the readout without learning a graph representation.** All heads use the same propagated feature family. A closure fraction above one is possible because the envelope is an observed model menu, not an optimum. Cornell’s reported closure fraction is 1.50.

| Task      | Linear | MLP-64 | Envelope | Interpretation at this budget          |
|-----------|--------|--------|----------|----------------------------------------|
| Cornell   | .5946  | .7027  | .6667    | Gap closes; quadratic probe only .4865 |
| Texas     | .7027  | .8378  | .8468    | Reported closure fraction .94          |
| Wisconsin | .7451  | .7647  | .7451    | Closes within reported spread          |

Across eleven held-out node tasks, the linear probe’s best value lies within the envelope spread on six; on Actor it exceeds the envelope by 2.3 points. The MLP closes only .090 of the remaining squirrel gap and .108 on BlogCatalog. Ten-seed BlogCatalog results remain 2.5 points below the envelope band throughout.

Two post-hoc studies explain, but do not rescore, frozen misses. WikiCS has frozen nonlinear gain .0121, above a .01 bar; the ten-seed mean is .0098 with 95% interval [.0076, .0120], and nine of ten seeds fall inside the envelope band. Coauthor Physics gives  $.9722 \pm .0004$  against a menu band of  $.9714 \pm .0003$ ; nine of ten seeds exceed the upper bar .9717. The latter is a reproducible exceedance of the finite menu, not an impossibility for that menu’s architectures under every training procedure.

On PascalVOC-SP (Dwivedi et al., 2022), the linear probe goes from  $R(0) = .0446$  to  $R(8) = .0455$  in macro-F1, with  $R(8) - R(1) = .0006$ . The MLP reaches .0581 and the quadratic probe .0627. The MLP improvement over the linear best is .0126, missing the frozen .05 prediction. These measurements identify unsuccessful probe families, not a proof of irreducible function-class difficulty.

#### E.5 TRANSFORMER BREADTH, REPORTED ARM BY ARM

Tables 22 and 23 distinguish the cross-system comparison from the four-variant sweep. The training-anchor row makes the two failure modes visible without inferring them from test accuracy alone.

Table 22: **Breadth-study comparison.** Entries are accuracy. The WikiCS generic envelope has reported spread  $\pm .0075$ . “Best GT” is wide on the FSM cells and deep on WikiCS. Missing entries are not inferred.

| Cell                     | Chance | Probe | Generic | Best GT | Typed | Exact |
|--------------------------|--------|-------|---------|---------|-------|-------|
| FSM transductive, ER = 5 | .1251  | .2027 | .2551   | .1965   | .6308 | .8149 |
| FSM inductive, ER = 245  | .1250  | —     | .2891   | .1822   | .7804 | .7831 |
| WikiCS                   | —      | .7852 | .7985   | .7555   | —     | —     |

Table 23: **Four variants beyond the registered GPS arm.** Each variant uses three seeds. The registered GPS arm reaches .7527 on WikiCS. Only two of four variants show the predicted perfect-anchor-fit/poor-test signature, so the prediction of at least three is a miss.

| Measurement                | Deep  | Wide  | PE-16  | Attention-only |
|----------------------------|-------|-------|--------|----------------|
| FSM test, ER = 5           | .1706 | .1965 | .1648  | .1510          |
| FSM test, ER = 245         | .1551 | .1822 | .1645  | .1424          |
| Training anchors, ER = 245 | .7253 | .7942 | 1.0000 | 1.0000         |
| WikiCS test                | .7555 | .7424 | .7395  | .6841          |

#### E.6 THE ACCESS-NORMALIZATION FAILURE

Table 24: **No universal per-entry knee in the registered window.** The study uses fifteen rungs and two seeds per rung. The knee is the first reported access level at which performance reaches at least .9 of the exact reference. All lie below the preregistered window [.5, 4]; the smallest family is at ceiling from the first rung. Pooled Spearman .326 fails the .6 criterion.

| $(C, S)$ | $D$ | Trace-length range | ER range   | Knee ER/ $D$ | Max spread |
|----------|-----|--------------------|------------|--------------|------------|
| (4, 2)   | 8   | 600–12,000         | 1–25       | $\leq .125$  | .0191      |
| (8, 4)   | 32  | 2,500–50,000       | 4–130.5    | $\simeq .39$ | .0468      |
| (16, 4)  | 64  | 5,000–100,000      | 12.5–243.5 | $\simeq .38$ | .0083      |

The earlier positive observation is narrower: measured FSM transductive and inductive cells order together by ER (Spearman 1.0), and the matched-access LDPC difference is .0088. A count that aligns a limited sweep need not identify all the supervision a solver can extract. Nonadjacent anchors constrain a small transition table even when the adjacent-pair count is one. The failed normalization is therefore not restored by renaming the count.

## E.7 THE OBSERVATION-LIMITED BRANCH

Epidemic studies force a third distinction alongside representation and learning access: the observation process can limit even a principled inference procedure. The decomposition study scores two of three predictions correctly and reports dynamic message passing restored to twice the probe floor under the tested observation change. A proposed three-feature ER collapse fails, including its power follow-up; a capacity test scores zero of two. Fixed rules on informative statistics outperform learned rules at low supervision in those experiments. The experimental record supports interacting limitations, not a numerical minimum of quantities on a common scale. We retain the bottleneck account as an organizing principle only.

## F FROZEN PREDICTIONS, STRICT SCORES AND THE NEGATIVE RECORD

### F.1 CONTROLLED COMPARISONS AND FROZEN-RECIPE TRANSFER

The six contracts below were frozen before their corresponding confirmation analyses or fits, with the preparation chronology stated for the strong-baseline comparison. Predictions concern their specified thresholds, not whether an outcome can later be given a useful interpretation. Scores are retained under the literal rules.

| Study                           | True/scored | Brier | Scope carried with the result                                 |
|---------------------------------|-------------|-------|---------------------------------------------------------------|
| Matched queries and FSM access  | 7/7         | .0429 | Independent gadget draws; FSM questions are not Bayes-matched |
| Circuit alignment and factorial | 8/8         | .0431 | Supplied scaffold; repeated splits/seeds of one board         |
| Strong Minesweeper baseline     | 6/6         | .1008 | Reproduced GNN/ensemble menu; not matched BuNN                |
| Fresh social-graph panel        | 2/6         | .0688 | No typed superiority; system gains are generic/ensemble       |
| Block-resampling sensitivity    | 5/5         | .0827 | Conditional on predictions; graph blocks may be dependent     |
| Fresh patent-view transfer      | 5/5         | .1250 | Overlapping induced views; one citation-time source graph     |

This is 33 true predicates among 37, not an aggregate calibration certificate. Four low-probability social-panel success events do not occur; preservation and the ensemble sanity check do. Their low Brier loss is compatible with that anticipated outcome. The predicates share models, datasets and sometimes logical dependencies, so a small aggregate score does not establish calibrated probabilities.

**Audit identifiers and exact construction constants.** Source records identify these contracts as P106, P107, P108, P109, P110 and P111, in the row order above. The patent root key is SHA-256 of the ASCII string `P111|view{v}|{u}`, with decimal view and source-node indices. The views and masks precede every fit. The source graph digest and view digests are retained in the numerical provenance data. A unique-edge-count assertion was corrected before any patent fit; membership, masks and sampling rules did not change.

**Literal endpoints and exposure.** The strong-baseline and patent-view endpoints require both candidate value over controls and acceptance, rather than counting a baseline return as an engine gain. The social panel’s positive rule requires at least four of six wins over both controls and two typed acceptances; its practically-null rule requires zero typed acceptances. Each dataset has no control win but at least one acceptance, so the frozen label is unresolved. The actual negative control comparison remains explicit.

The historical FSM shared-array study trained on observed-anchor positions and reported a test-best radius. It scores its original predicates under that protocol. The disjoint observation/supervision experiment reported in the main text uses a distinct contract and greater total label access; it does not rescore that historical forecast or isolate a single cause of the numerical change. The strong-baseline preparation logged held-out metrics, while decisions used selection values as detailed in [Appendix B.7](#). Forecast hashes alone do not remove those distinctions.

## F.2 THE 54-PREDICTION PROSPECTIVE PANEL

The list contains 20 tasks. LastFMAsia and its epidemic variant were unavailable at source before any probe ran, leaving 18 evaluated tasks. Prediction dependence is represented by 19 scorer-defined task clusters; those keys need not coincide one-to-one with the displayed task names.

| Family                  | Frozen | Correct | Frozen content and observed limitation                                                                                                                                          |
|-------------------------|--------|---------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fault localization      | 12     | 12      | Unique-signature accuracy 1.0 on four held-out circuits; typed at least generic, with a twofold result on c 6288. Exactness is coverage-qualified.                              |
| Saturation containment  | 33     | 32      | Eleven tasks, three related predictions each. Accepted gain at most .5 points; final soft gate below .15. Observed gate at most .045; one spread-band violation remains a miss. |
| Access escape           | 2      | 2       | Ensemble-access escape on new (4, 8) codes and (3, 6) at $n = 4800$ .                                                                                                           |
| Plateau immobility      | 4      | 2       | Affine obstruction survives. Generic learned-arm behavior is not universally immobile, particularly at high supervision.                                                        |
| Epidemic source finding | 2      | 1       | A miss separates observation/reference limits from learner-access limits.                                                                                                       |
| Size extrapolation      | 1      | 0       | Refuted as stated.                                                                                                                                                              |
| Total                   | 54     | 49      | Accuracy .907; Brier .1139.                                                                                                                                                     |

Equal weighting of the six families yields accuracy .662 and Brier .203. A  $10^4$ -draw task-cluster bootstrap gives interval [.804, .982] for prediction-weighted accuracy. Leaving out containment gives the lowest leave-one-family-out accuracy, .810. Fault localization and containment together score 44/45; the four extrapolation families score 5/9. There is no held-out node task predicted to be nonsaturated in this panel, a coverage gap rather than a successful test of that regime.

## F.3 THREE PREREGISTRATIONS AND A STRICT CORRECTION

Preregistration A covers P 1–P 13: nonlinear probes, isolated dispatch, transformer breadth, the query pilot and the LDPC correction. B covers P 14–P 18: the Minesweeper factorial and wrong-law refusal. Its membership axis was amended from a degenerate identity condition to random membership before any run; that amendment belongs in the immutable record. C covers P 19–P 25: the clean shared-evidence query intervention and second-family factorial.

The shipped scorer for P 17 accepted any worst typed condition with randomized membership. The frozen statement named the *unanchored* random-membership typed condition. The observed worst cell was anchored. Literal scoring therefore changes P 17 from pass to miss. The strict standing score for P 1–P 18 is 12/18, Brier .2106; the preserved lenient result is 13/18, Brier .1828. P 19–P 25 score 6/7, Brier .1471. Together, the strengthening study has eighteen passes and seven misses, listed below. Post-hoc explanations do not change these scores.

Table 25: **Every strict miss in the strengthening study.** The right column describes later interpretation, not a revised verdict.

| ID  | Frozen bar and outcome                                                             | Follow-up and retained implication                                                                                                                                 |
|-----|------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| P 2 | WikiCS nonlinear gain at most .01; observed .0121.                                 | Ten-seed mean .0098, interval [.0076, .0120], nine of ten seeds within the envelope band. The frozen single-seed forecast still misses.                            |
| P 3 | PascalVOC-SP nonlinear gain at least .05; observed .0126.                          | No rescoring or resolving follow-up. The inexpensive radius and readout probes do not close this task.                                                             |
| P 4 | Probe no higher than Coauthor Physics envelope plus spread .9717; observed .9722.  | Ten-seed mean $.9722 \pm .0004$ , with nine of ten above the bar. This finite model envelope is not a ceiling.                                                     |
| P 5 | At least ten of eleven dispatch decisions persist under isolation; observed eight. | Withdrawn acceptances are Actor, squirrel and Reddit; all had zero original test gain. The correction exposes validation reuse, but does not satisfy the forecast. |

## Strict misses, continued

| ID   | Frozen bar and outcome                                                                                   | Follow-up and retained implication                                                                                                      |
|------|----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| P 9  | At least three of four new transformers show perfect anchor fit and poor test accuracy; observed two.    | Two variants memorize; two also fail to fit. A subsequent fitting check does not retroactively make the single-mode prediction correct. |
| P 17 | Worst typed cell predicted to be random membership without anchors; anchored random membership is worse. | Strict miss replaces a lenient pass. A newly frozen directional claim in FSM (P 25) succeeds: $.1473 \leq .1483$ .                      |
| P 24 | FSM anchor effect must reach the specified twofold bar .0382; observed .0378.                            | Miss by .0004. The larger membership/operator interaction does not depend on passing this separate threshold.                           |

## F.4 LATER BOUNDED STUDIES: KEEP OUTCOMES SEPARATE

| Study                                      | Correct/scored | Brier | What the score does not hide                                                        |
|--------------------------------------------|----------------|-------|-------------------------------------------------------------------------------------|
| Independent-board count reuse              | 6/6            | .060  | One source-fit regime and common accepted support                                   |
| Integrated ten-task system                 | 5/6            | .2204 | Strict Minesweeper .990903 misses the frozen .992 bar                               |
| Arxiv replication and perception checks    | 4/6            | .1021 | Perception upgrade and Computers gate miss; one conditional non-run                 |
| Independent-GNN control and transfer gates | 3/4            | .1344 | Both breadth gates fail; two conditional non-runs                                   |
| Local node-to-joint export                 | 0/4            | .2344 | No successful integrated joint reuse                                                |
| Coherent forest prototype                  | 2/4            | .0963 | Preservation/boundary hits, not useful dependence                                   |
| Stable local-joint prototype               | 3/5            | .1350 | Correctness, preservation and one acceptance hit; both useful-query predicates miss |

Conditional non-runs are not counted as wins or losses; failed prerequisites remain scored. The studies share data, mechanisms and development history and are not pooled into one program-wide success rate. In particular, the historical 111-forecast aggregate has unresolved per-metric inclusion masks and is omitted. This record supports procedural traceability, not a claim of well-calibrated probabilities.

The earlier representation census has literal scores 1/11 and 4/13, with a failure falsifier fired; its fold-basis reinterpretation is post-hoc. The Roman direction study scores 5/8 literally, not its post-hoc 7/8 interpretation; one frozen item had no probability. The arxiv direction study scores 7/7 on its registered random splits, with official-split replication descriptive. The event-query study scores 1/6; same-board learned-law reuse scores 5/7 and does not prove independent-instance transfer. The filtered direction study scores 3/6, including one vacuous hit caused by a negative reference gain; its matched-recipe correction is exploratory. These records are retained rather than silently counting useful measured effects as successful forecasts.

## F.5 TESTS THAT COULD OVERTURN THE EMPIRICAL ACCOUNT

The registered challenges are: a newly preregistered L 0 arm, at the declared low-ER budget in a measured family, closing at least half the reference headroom; a probe-tight task ( $\Delta_R < .05$ , envelope gap below .01) on which a new architecture beats the menu by more than twice its registered spread; a matched-ER access-mode effect exceeding .05 of exact-reference performance; more than 30% regime mislabeling in a held-out family under the frozen protocol; or a new factorial in which typed operators benefit randomized membership as much as true membership.

These challenges concern the empirical regularities, not the proved affine theorem. Conversely, exceeding a genuinely exact *population* optimum on its declared information and identifiability domain would signal inconsistent assumptions, implementation or evaluation, rather than greater-than-Bayes inference. A finite test estimate above a reference is not, by itself, proof of leakage.

## G ORIGINAL CFN, TABLE RECOVERY AND HISTORICAL DISPATCH

### G.1 CARDINALITY FACTORS AND LOGICAL ENTAILMENT

The original CFN learns tables indexed by an observed clue and the number of mines among its factor members. It is given the cardinality family and true clue membership, not the completed numerical table. The ten official split means are reported in Table 26. These are different fits and a different branch from the integrated results in Table 5.

Table 26: **Original CFN branch, not the integrated panel.** Minesweeper AUC. Published references are separate from the matched circuit intervention and the project’s own fitted arms. Historical dispersion terms for CFN are not assigned an unsupported confidence-interval interpretation.

| Model                      | Supplied information / supervision                        | AUC               |
|----------------------------|-----------------------------------------------------------|-------------------|
| Unary                      | No relational factor                                      | .5115             |
| Affine pairwise            | Pairwise factor class                                     | .8934 $\pm$ .0059 |
| Published tuned GCN*       | Training labels; no label inputs (Luo et al., 2024)       | .9786 $\pm$ .0024 |
| Published tuned GraphSAGE* | Training labels; no label inputs (Luo et al., 2024)       | .9777 $\pm$ .0062 |
| Published BuNN             | Training labels; no label inputs (Bamberger et al., 2025) | .9899 $\pm$ .0016 |
| CFN                        | Count family; training labels, no label inputs            | .9906 $\pm$ .0008 |
| CFN                        | Count family; training labels also as inputs              | .9998 $\pm$ .0001 |

The BuNN reference is the published ten-seed mean and standard deviation. The CFN no-label-input mean exceeds it by .0007 AUC, or .07 percentage points; this numerical comparison is not a paired significance test. CFN additionally receives the count family and clue membership. These references situate its benchmark performance; the controlled recompilation experiment and integrated panel separately identify circuit value and matched system gains.

Under the original benchmark’s observed clues and revealed training labels, sound constraint propagation determines a reported 97.2% of test nodes, with zero observed contradictions among the determined nodes. Without training anchors the propagation study determines 57.0% of all nodes. These denominators differ: the 97.2% figure is neither label-free coverage nor a certificate for the CFN’s predictions. It is not an AUC ceiling. Later posterior-thresholded coverage at probabilities below .01 or above .99 is a separate confidence statistic, not logical entailment.

For equality clues  $c = 0, \dots, 4$ , covering 98.7% of clued nodes, every exercised table row has its maximum at the true count. The clipped  $\geq 5$  bucket recovers the corresponding admissible support. Unexercised entries remain unidentifiable. Figure 8 summarizes reported maxima/support rather than reconstructed probabilities.

A different follow-up on Questions and Amazon-ratings finds stable signs in extracted factor rules. On Questions, a two-parameter rule retains at least half the engine gain on nine of ten splits; Amazon’s extracted rule is standalone-predictive against its null. These small-effect findings are not Minesweeper count-table identification and are not used to enlarge its claim.

### G.2 KNOWN-LAW LOCAL INFERENCE

Supplying the numerical observation law removes the table-learning problem. Exact variable elimination in radius-three and radius-four local models gives reported mean AUC about .9957 and .9982, without gradient fitting. The local window and width-limited fallback are part of the model: exact local marginals are not guaranteed to equal whole-board marginals. The integrated node evaluation also records a supplied-count loopy-BP arm; its local contractions are exact but its global inference is approximate. Neither supplied-law branch earns credit for a learned-only system comparison.

### G.3 THE EARLIER ISOLATED REPLACEMENT RULE

The historical dispatch study, separate from the current mixture/threshold procedure, compares three saved candidate checkpoints with a saved perceiver on a held-out validation half. Ties refuse. Ten of eleven node tasks refuse, and Coauthor CS accepts with zero measured test gain. Eight of

eleven decisions persist after isolation, not the predicted ten. Actor, squirrel and Reddit lose prior acceptances that had yielded zero test gain. These observations do not justify reusing validation for both checkpoint selection and final acceptance.

Table 27: **Utility, not law purity.** Historical FSM deployment study. A wrong table is refused; an imperfect noise model can still improve on the baseline.

| Candidate              | Validation   | Deployed test | Decision       |
|------------------------|--------------|---------------|----------------|
| Perceiver              | .261         | .255          | Baseline       |
| True transition law    | .7833        | .7789         | Accept         |
| Wrong transition table | .2318        | .255          | Exact fallback |
| Wrong noise model      | Not reported | .6561         | Accept         |

The earlier soft-gated system’s weight measurements and aggregate gates are not a theorem about the current deployment rule. The IID accuracy lemma in [Appendix A](#) applies only to the argmax procedure it explicitly states.

## H REPRODUCIBILITY AND REPORTING CONVENTIONS

### H.1 POSITIONING THE TWO CONSTRUCTIVE DEMONSTRATIONS

[Table 28](#) compares the objects that methods learn and execute, rather than assigning broad capability checkmarks. The last two rows are distinct implementations. Having both in a research program does not show that one system unifies them.

Table 28: **Different learned objects support different computations.** Selected comparisons, not an exhaustive capability audit. No row implies that related work cannot be extended to other queries.

| Method                                                                     | Supplied structure                                      | Learned or selected object                | Execution / evaluation                                             |
|----------------------------------------------------------------------------|---------------------------------------------------------|-------------------------------------------|--------------------------------------------------------------------|
| GDC ( <a href="#">Gasteiger et al., 2019b</a> )                            | Graph and diffusion construction                        | Downstream predictor                      | Prediction on transformed connectivity                             |
| Neural relational inference ( <a href="#">Kipf et al., 2018</a> )          | Interaction-model family                                | Latent edges and dynamics                 | Neural message-passing dynamics                                    |
| Discrete structure learning ( <a href="#">Franceschi et al., 2019</a> )    | Graph-learning hypothesis class                         | Structure and GNN parameters              | Bilevel learning for prediction                                    |
| MetaGL ( <a href="#">Park et al., 2023</a> )                               | Candidate model set and graph descriptors               | Model selection rule                      | Select a graph learner                                             |
| FGNN ( <a href="#">Zhang et al., 2020</a> )                                | Factor-graph representation                             | Neural variable/factor message updates    | Higher-order prediction and approximate inference                  |
| Learned factor graphs ( <a href="#">Shlezinger et al., 2022</a> )          | Stationary Markov factorization                         | Local conditional models                  | Reuse in sum-product and Viterbi inference across sequence lengths |
| Neural-enhanced BP ( <a href="#">Garcia Satorras &amp; Welling, 2021</a> ) | Factor graph and BP computation                         | Neural message corrections                | Hybrid inference, including channel decoding                       |
| Integrated TSGL here                                                       | Graph, factor library, bounded membership policy        | Perceiver, local tables, mixture proposal | Node prediction with held-out acceptance                           |
| Count-family compiler here                                                 | Count family, grid membership, local inference language | Source observation table                  | Frozen-table reuse for three event families                        |

These methods, together with differentiable and Gaussian belief-propagation approaches ([Opipari et al., 2021](#); [Cui et al., 2024](#); [Nabarro et al., 2024](#)), establish learned factors and hybrid inference as useful ingredients. Here the experiments test a specific decomposition of graph-learning tasks and its consequences. The compiler’s contribution is evidence of frozen-table transfer and dependence value

against its own-marginal controls in a declared count family; it is not the first use of a learned local model inside probabilistic inference.

## H.2 CODE AVAILABILITY AND NUMERICAL REPRODUCTION

**Code available on request.** The corresponding author can provide experiment drivers, configurations, frozen sampling records, masks and retained predictions. The manuscript source includes the numerical inputs for its tables and vector figures; compiling it does not require training models or reading the original experiment storage.

The paper-source checks independently enumerate the small ranking constructions and verify selected displayed quantities. Such checks establish arithmetic and implementation identities, not the correctness of every scientific interpretation. Full trained-model reproduction requires the cited public datasets, the declared access/split conventions, dependencies and suitable compute.

**What is retained.** The circuit-alignment and integrated-system comparisons retain per-node predictions and exact masks. Independent replay checks their metrics and acceptance decisions. The fitted binary-query and FSM-readout studies retain scalar curves and selected configurations, but not their fitted weights or score arrays. The replicated Minesweeper factorial retains scalar results, not checkpoints or timings. Those experiments require refitting for a prediction-level replay. The historical circuit experiment also lacks the predictions needed to replace its sequential-rank AUC with tie-aware scoring; its values are kept as a separate record, not declared numerically invariant.

The count-family reuse path retains the frozen learned tables and evaluates supplied-law, learned, untrained and own-product arms on common supported target windows. A model that answers exactly inside that finite local representation is not thereby an exact full-board model. Some historical diagnostic/checkpoint collections and the original .908 diagnostic scorer are unavailable; the reported replacement audit is a different calculation.

**Reproduction order and access.** Partition training, selection, acceptance and test masks before fitting the integrated node system. Fit the retained perceiver and candidates under the declared label access; select checkpoints, candidates and mixture weights without acceptance labels. Lock predictions, apply the acceptance rule, and then score test outputs. Report baseline, standalone specialist, locked candidate and deployed system separately. Across views of the same graph, do not reinterpret repeated nodes or splits as independent populations.

For count-family reuse, freeze the source table and prior before target-board generation. Reconstruct the declared count family, grid membership and anchors; execute all arms on common supported local windows. Preserve abstentions and support counts as well as answers. For the diagnostic studies, retain each probe version and its actual selection rule; a test-best historical curve is not a validation-selected deployment.

## H.3 METRICS, UNCERTAINTY AND PROVENANCE

AUC, accuracy, macro-F 1 and MRR retain their native meanings. Percentage points are 100 times a difference on the unit scale. Split means on one graph are not independent populations; seed variation is not a substitute for independent-domain replication. The independent-board bootstrap uses boards as units. Other paired-split intervals describe the tested split variability only.

The original generic-menu spread is its reported maximum-minus-minimum; some historical  $\pm$  terms lack a resolved SD/SE/CI definition and are not relabeled here. Every current table states its aggregation explicitly. The spatial/graph block bootstrap is a sensitivity analysis conditional on saved predictions, not an independent-graph confidence statement. Neither an exact fallback nor a high-confidence prediction supplies an unconditional guarantee about accepted specialists.

Only the named frozen studies are prospective. Development choices, corrected scorers, post-hoc seed analyses, the locked ungated ensemble comparison and the no-law mixture diagnostic retain their respective statuses. The unresolved program-wide forecast aggregation is excluded rather than presented as a calibration credential.