TASK SIMILARITY PREDICTS IN-CONTEXT TRANSFER ON GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Whether a graph foundation model will learn a new task in context cannot currently be predicted before training. The question is open because alignment theory for in-context learning is derived for linear models and does not reach graphs, and because standard graph few-shot evaluations conflate in-context inference with structural shortcuts such as label propagation. We construct a cross-graph rule-induction testbed that closes these shortcut channels, with no edge between demonstrations and queries and with the label polarity randomized per episode. On this testbed a model's in-context gain is governed by a single measurable scalar, the polarity-symmetrized label agreement between the training task and the test task. A Bayesian posterior-concentration argument predicts the full few-shot gain surface with no free parameters, and predicts that adding graph families or rules brings no gain unless it changes the agreement statistics. The law holds on synthetic rules, on five real graph datasets, and on a frozen, externally pretrained graph foundation model (GraphPFN). In-context transfer on graphs can therefore be forecast from labeled examples alone, without model access or any training, and the same measurement can guide pretraining task design.

1 INTRODUCTION

In-context learning (ICL) enables a frozen model to solve a new task from a small number of demonstrations in its input without gradient updates. On graph data, PRODIGY (Huang et al., 2023) and GraphPFN (Eremeev et al., 2025), a prior-data fitted network (Müller et al., 2021) in the tradition of TabPFN (Hollmann et al., 2022; 2025), extend this paradigm to node-level prediction, with the goal of adapting a pretrained model to new graphs and tasks from examples alone. Recent work on in-context regression shows that the relation between the pretraining tasks and the test task governs in-context generalization (Lu et al., 2025; Letey et al., 2026; Goddard et al., 2025; Wang et al., 2025). The theory is derived for linear regression in solvable models, the experiments train transformers on regression tasks or probe language models on synthetic tasks, and none of this work covers graph data. We therefore ask the following question. **Can in-context transfer on graphs be forecast before training or deployment, from labeled examples of the source and target tasks alone, without access to the model?**

Addressing this question first requires separating in-context inference from the shortcut channels that transductive graph evaluation leaves open. When demonstrations and queries share edges in one connected graph, a model can exceed chance accuracy without using the support labels. We construct a *cross-graph rule-induction* testbed (Figure 1) that closes each of these channels by construction; Section 2 states the four isolation requirements and the corresponding control experiments. On this testbed we train four model classes, spanning three ICL mechanisms, from scratch, and measure their transfer to held-out tasks.

Across the four architectures and three ICL mechanisms, the in-context gain Δ_{ICL} is predicted by a single scalar, the polarity-symmetrized train–test label agreement s, with Pearson correlations of $r = +0.94$ to $+0.98$ (Section 2.2). s is computed from labeled examples of the two tasks and requires neither training nor model access. On the same runs it outperforms seven alternative predictors, including transferability estimators that require the trained checkpoint.

Section 3 derives this law from a Bayesian analysis of the meta-training objective. The analysis predicts the full few-shot surface $\Delta_{\mathrm{ICL}}(K, s)$, as a function of the demonstration count K and the

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

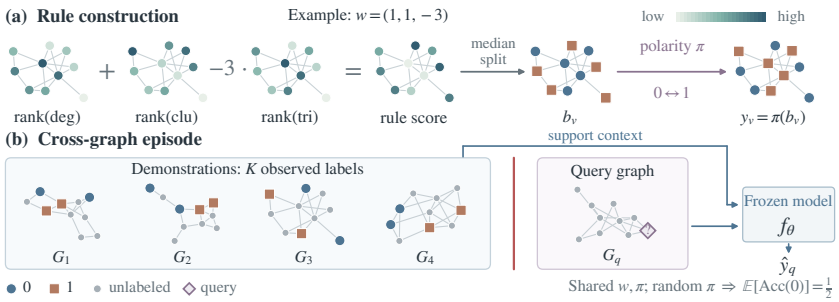

Figure 1: **The cross-graph rule-induction testbed.** (a) A task is a rule w over within-graph ranks of node statistics, here $w = (1, 1, -3)$ over degree, clustering, and triangle count. The weighted score is split at the median of the graph, which gives balanced labels b_v, and a per-episode polarity π maps them to $y_v = \pi(b_v)$. (b) An episode labels four demonstration graphs G_1 to G_4 and one query graph G_q with the same w and π; the model observes K support labels on the demonstration graphs and predicts the query label with frozen parameters. The query graph shares no edge with any demonstration, which excludes label propagation, and the random polarity fixes zero-shot accuracy at chance, so any gain at $K > 0$ requires reading the support labels.

similarity s, with no free parameters. The analysis implies that the gain depends on the training pool only through its agreement statistics, so adding graph families or rules brings no gain unless it changes those statistics. Sweeps over graph families and rule count find no gain (Section 4.1). The analysis also specifies when diversity does increase the gain, namely when the members' node-wise majority vote agrees with the target more than any member does. In a pre-registered experiment, a rule pool with complementary member errors outperforms its best individual member by an amount consistent with the prediction.

The same law holds outside the synthetic testbed. On five real graph datasets, s forecasts transfer out of sample (Section 4.2). The frozen, publicly released GraphPFN checkpoint (Eremeev et al., 2025), which we evaluate without fine-tuning, follows the law at its native scale with efficiencies of 0.88 to 0.97 relative to the reference learner (Section 4.3). The conclusions therefore do not depend on our training procedure or model choices. Related work is discussed in App. A.

2 A TESTBED AND AN EMPIRICAL LAW

The testbed is designed to satisfy four requirements. Each closes one shortcut channel that transductive graph evaluation leaves open: (1) no edges between the demonstration graphs and the query graph, which excludes label propagation; (2) no labels on the query graph, which excludes homophily-based prediction; (3) zero-shot accuracy equal to chance, which excludes any label convention stored in the weights; and (4) a query label that is not predictable from query-side features alone. Under these four conditions, above-chance accuracy can arise only if the model conditions its prediction on the support labels. Section 2.1 describes the construction and Section 2.2 reports the resulting empirical regularity.

2.1 THE CROSS-GRAPH RULE-INDUCTION TESTBED

Rule construction. Controlled ICL studies (Garg et al., 2022; Raventós et al., 2023; Letey et al., 2026) define a task as a latent linear map that the model must infer from demonstrations. The graph analogue requires a rule that is defined on any graph, is consistent across graphs of different sizes and degree distributions, produces balanced labels, and belongs to an enumerable family, so that train–test similarity can be controlled. A task is a rule w that labels every node of a graph (Figure 1a). For each node, the rule computes three structural statistics (degree, clustering coefficient, and triangle count). Each statistic is ranked within its graph and the rank is scaled to $[0, 1]$; ranking makes a single weight vector applicable to graphs of different sizes and degree scales. The rule then forms the weighted score of the ranks and thresholds it at the median of the graph, so that the two classes are balanced on every graph. A polarity π, drawn uniformly at random for each episode, is then applied to all labels in the episode. Because π is independent of the unlabeled episode data, any predictor

that does not read the support labels has expected accuracy $1/2$ (Corollary 1). Two further rule families use the same rank, weight, and median-split pipeline with different inputs. **Feature** rules operate on node-feature coordinates, with synthetic features drawn as i.i.d. 8-dimensional noise. **Mixed** rules rank structural statistics and feature coordinates jointly and weight them together.

Transfer pairs. Training rules are integer weight vectors over the statistic basis with all weights nonzero. Because this family is enumerable, training pools can be constructed at prescribed similarity levels, and the label agreement between any training rule and any evaluation task can be measured on a common graph sample. Transfer is always evaluated on two fixed held-out tasks. The *majority* task labels each node by the balanced majority of the three statistics and is correlated with each of them. The *selector* task labels by a single statistic chosen by a latent variable, so its agreement with a training rule varies across the three latents; results are averaged over the three choices. With the evaluation tasks fixed, varying the training rule changes only the train–test similarity.

Episode protocol. An episode consists of $m{=}4$ demonstration graphs, each sampled independently from a prior over four graph families, Erdős–Rényi (ER) (Erdős & Rényi, 1959), stochastic block model (SBM) (Holland et al., 1983), Barabási–Albert (BA) (Barabási & Albert, 1999), and Watts–Strogatz (WS) (Watts & Strogatz, 1998), with 32–64 nodes and mean degree 6. All nodes of all graphs in the episode are labeled by one rule w under one polarity π. The model observes K of these labels as support pairs distributed over the demonstration graphs. Within a graph the support nodes alternate between the two classes from a random starting class, so that the identity of a support node carries no information about π (Corollary 1). The model predicts the label of one query node v_q on an additional, independently sampled graph. Formally, an episode is $E = \big(\{(G_i, S_i)\}_{i=1}^{m}, G_q, v_q; w, \pi\big)$ with support pairs $S_i = \{(v, \pi(w(G_i))_v)\}$, $K = \sum_i |S_i|$, and target $y_q = \pi(w(G_q))_{v_q}$. We write E_{obs} for the observed part of E, that is, E without (w, π). Models are meta-trained on the episode stream by minimizing $\mathbb{E}_E\, \mathbb{E}_K[-\log p_\theta(y_q \mid E_{\mathrm{obs}})]$ over θ, where K is drawn for each episode from $\{0, 1, 4, 8, 16\}$. At evaluation, θ is frozen and the model is presented with 600 new episodes of a held-out task at each support size. The quantity of interest is the *in-context gain* $\Delta_{\mathrm{ICL}} = \mathrm{Acc}(K{=}16) - \mathrm{Acc}(0)$. Both terms are fixed by the design. $\mathrm{Acc}(0)$ equals chance in expectation by Corollary 1, and requirements (1), (2), and (4) close the remaining channels, so a positive gain is attributable to the support labels.

Model classes. We evaluate four model classes that cover the principal mechanism families of current graph few-shot systems, each trained from scratch on identical episode streams. Model A is a GraphSAGE encoder (Hamilton et al., 2017) with a prototype metric readout (Snell et al., 2017). Model B uses the same encoder with a PRODIGY-style metagraph readout (Huang et al., 2023), so that the pair isolates the effect of the readout. Model C1 is a set transformer (Lee et al., 2019) over support and query tokens, and Model C2 is a full transformer (Vaswani et al., 2017) over all episode nodes; neither uses message passing, and both rely on PFN-style attention (Hollmann et al., 2022). The four models span three ICL mechanisms, so that the results are not specific to one architecture. All four models receive the same node input, the node features concatenated with per-graph standardized degree, clustering coefficient, and triangle count. Architecture and training details are given in App. G. Each isolation requirement is tested by a control experiment (a label-propagation baseline, a polarity-switch control with randomization on and off, a query-feature probe, support-label permutation, wrong-rule context, and a support-order control). Definitions are given in App. G; results are reported in Section 2.2 and App. C.1.

Similarity measurement. To isolate the effect of the training–evaluation relation on the gain, we use a single-rule design. Each model is trained from scratch on the episode stream of one training rule and evaluated on the two held-out tasks. (A training *pool* of T rules generalizes this design. The results below use $T{=}1$; Section 3 treats $T > 1$.) The graph families, m, the distribution of the training K, the training budget, and the 600-episode evaluation are identical across runs, so that runs differ only in the similarity between the training rule and the evaluation tasks. Similarity is defined as follows. For a training rule r and an evaluation task e, the per-node label agreement on a graph G is $a_G(r, e) = |V_G|^{-1} \sum_v \mathbf{1}\{r(G)_v = e(G)_v\}$, and its polarity symmetrization is $\sigma_G(r, e) = \max\big(a_G(r, e),\, 1 - a_G(r, e)\big)$. Symmetrization is required because, under polarity randomization, a rule and its label complement induce the same distribution of episodes. The similarity $s(r)$ is the mean of σ_G over a fixed sample of 1,250 graphs from the training families, over both evaluation

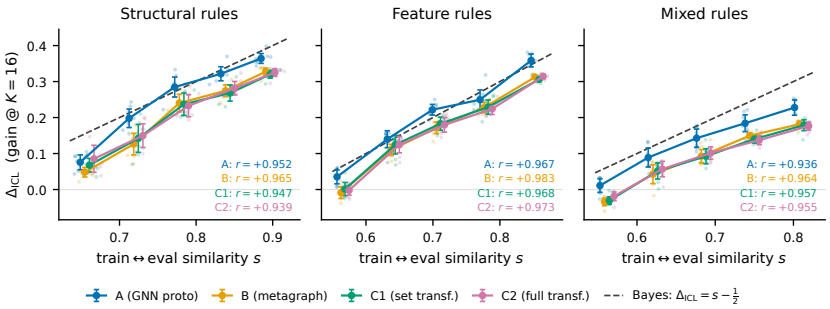

Figure 2: **In-context gain tracks measured similarity in every cell.** The gain Δ_{ICL} (accuracy at K=16 minus measured zero-shot) against measured train–test similarity s, for the single-rule training runs per model per ladder (40 rules, App. C.1), with light points for individual rules and markers for the per-level mean ± 95% CI. The dashed line is the zero-parameter Bayesian prediction $\Delta_{\mathrm{ICL}} = s - \frac{1}{2}$, derived in Section 3. Each panel prints its per-model Pearson r.

tasks, and over the three selector latents. Computing s requires labeled examples of the two tasks on common inputs, here the fixed graph sample, and requires neither training nor access to a model. The per-graph symmetrization differs from the global-polarity agreement of Section 3 by a folding term that App. D.2 measures. On this axis we construct a ladder of forty training rules, eight at each of five similarity levels. The forty rules of a ladder span 27 (structural), 22 (feature), and 27 (mixed) distinct weight directions. A rule and its negation label every graph as complements up to the median node, so under polarity randomization they induce the same episode distribution up to that node and serve as replicate runs. The three ladders (structural, feature, and mixed) crossed with the four model classes give $3 \times 4 = 12$ cells of 40 runs each.

2.2 The in-context gain tracks similarity

In all twelve cells, the Pearson correlation between Δ_{ICL} and s over the single-rule runs is between $+0.94$ and $+0.98$ (Figure 2). The control experiments confirm that the gain arises from in-context inference. Zero-shot accuracy is at chance (0.50 on Model A; 0.533 to 0.535 on B/C1/C2, the K=0 argmax-tie value of Remark 1), label propagation is at chance, and permuting the support labels or supplying a wrong-rule context removes the gain. The dashed line in Figure 2 is the zero-parameter prediction $\Delta_{\mathrm{ICL}} = s - \frac{1}{2}$. Model A follows this line closely. The other models show the same dependence on s with a smaller slope, which Section 3 models as an efficiency coefficient η.

Partial correlations controlling for final training loss are nearly unchanged in every cell (Table 2), so the correlation is not explained by differences in trainability. On correct-class probability, the correlations are $+0.97$ to $+0.99$ across all cells, and all results are replicated on log-margin and negative log-likelihood, which rules out accuracy-threshold artifacts (Schaeffer et al., 2023) (App. C.1). The correlation is reproduced under a T=8 pool-similarity sweep ($r = +0.921$, Model A), under end-to-end retraining with three seeds ($r = +0.978$ on seed-averaged Δ_{ICL}), and across hidden widths from 32 to 256, a $58\times$ to $60\times$ range in parameter count (App. C.1, App. C.2, and App. C.4).

On the same runs, we compare s with seven alternative predictors (Table 3). s attains the highest correlation and the highest leave-one-out Q^2 for every model ($r = +0.94$–0.96, $Q^2 = 0.87$–0.92), where Q^2 is the out-of-sample R^2 computed from the prediction residuals. Raw (unsymmetrized) agreement reaches only $r = +0.13$–0.18, so polarity symmetrization is necessary. LogME (You et al., 2021) is the strongest alternative ($r = +0.73$–0.92) but requires the trained checkpoint. The dashed line in Figure 2 is not fitted to the data; Section 3 derives it and states further predictions beyond the asymptote.

3 A Bayesian account of cross-task graph in-context transfer

Section 2.2 shows that the label agreement s between the training task and the evaluation task predicts the in-context gain. This section derives that dependence for a Bayesian reference learner.

On the training distribution the population minimizer of the meta-training objective is the posterior predictive over the training hypotheses (Theorem 1). Support sets that are inconsistent with every training hypothesis never occur under training, and the objective does not constrain the learner on them. The reference learner extends the posterior decision to such support sets by scoring each hypothesis through an independent label-symmetric noise channel, an assumption that the objective does not supply. For a single training rule the resulting classification decision is invariant to the channel noise level, and the accuracy it attains is a function of K and s with no free parameter. A trained network is compared with this accuracy through an efficiency coefficient η. We treat binary classification with a single training rule ($T{=}1$) first and extend to multi-rule pools in Section 3.3, where the noise level enters through the weight β.

3.1 The reference learner

Suppose meta-training uses a single fixed training rule r. In each episode the rule is the same and the labels are flipped by an independent random polarity $\pi \in \{+, -\}$; under $\pi = -$ the labels follow the complement rule $\bar{r}$. For an optimal learner that knows the training distribution, the latent variable inferred from the demonstrations is therefore the episode polarity, and the hypothesis class is $\mathcal{H} = \{r, \bar{r}\}$. Meta-training minimizes the negative log-likelihood $\mathcal{L}(q) = \mathbb{E}\big[-\log q(y_q \mid E_{\mathrm{obs}})\big]$. Because cross-entropy is bounded below by the conditional entropy, $\mathcal{L}(q) \geq \mathbb{E}\, H\big(P(y_q \mid E_{\mathrm{obs}})\big)$ with equality if and only if $q(y_q \mid E_{\mathrm{obs}}) = P(y_q \mid E_{\mathrm{obs}})$, so the population minimizer of the meta-training objective is the posterior predictive distribution under the training prior (Müller et al., 2021).

Theorem 1 (Population-optimal learner under single-rule training)**.** *For a polarity-randomized meta-training distribution with $T{=}1$, the population minimizer of the meta-training objective is the posterior predictive over $\mathcal{H} = \{r, \bar{r}\}$. On the training distribution its prediction depends on the support labels only through the episode polarity, which any single support label determines.*

Now consider a held-out evaluation rule e. In general the test support set is consistent with neither hypothesis in $\mathcal{H}$, and the noiseless likelihood of the training distribution does not apply. The reference learner scores each hypothesis through an independent label-symmetric channel, under which hypothesis h assigns probability $1 - \varepsilon$ to a support label it predicts correctly and ε to one it predicts incorrectly ($0 < \varepsilon < 1/2$). If $h^+ = r$ matches m of the K support pairs, the complement hypothesis $h^- = \bar{r}$ matches $K - m$, and the log-likelihood ratio is $\log\big[p(S \mid h^+)/p(S \mid h^-)\big] = (2m - K)\log\frac{1-\varepsilon}{\varepsilon}$. Since $\log\frac{1-\varepsilon}{\varepsilon} > 0$, the decision depends only on the sign of $m - K/2$, whatever the value of ε; r is favored when $m > K/2$ and $\bar{r}$ when $m < K/2$. Polarity inference on a held-out task therefore reduces to a majority vote over the K support agreements.

3.2 The transfer law under finite context

Define the aligned hypothesis as $r^* = \arg\max_{h \in \{r, \bar{r}\}} P[h(X) = e(X)]$ and its global label agreement with the test rule as $s = P[r^*(X) = e(X)] \in [1/2, 1]$. The measured similarity of Section 2.1 symmetrizes per graph before averaging and is therefore at least this global-polarity agreement. On the structural ladder the difference is 0.026 at the lowest level and 0.000 at the highest (App. D.2). For the i-th support node, define the agreement indicator $A_i = \mathbf{1}\{r^*(X_i) = e(X_i)\}$. Under independent sampling, $A_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(s)$, and $M_K = \sum_{i=1}^{K} A_i \sim \mathrm{Binomial}(K, s)$. The probability that majority voting recovers the aligned polarity is $P_c(K, s) = P[M_K > K/2] + \frac{1}{2}\,P[M_K = K/2]$. If majority voting selects the aligned polarity, the query prediction agrees with the test task with probability s; if the opposite polarity is chosen, the accuracy is $1 - s$. Because query-node agreement is independent of the support votes, $\mathrm{Acc}(K) = P_c\, s + (1 - P_c)(1 - s) = \frac{1}{2} + (2P_c - 1)(s - \frac{1}{2})$. Subtracting $\mathrm{Acc}(0) = 1/2$ yields the following.

Theorem 2 (Finite-context similarity law)**.**

$$\Delta_{\mathrm{ICL}}(K, s) = \big(2P_c(K, s) - 1\big)\left(s - \frac{1}{2}\right).$$

The expression has no free parameter. The first factor $(2P_c - 1)$, the vote factor, is the probability advantage of identifying the polarity from K demonstrations. The second factor $(s - \frac{1}{2})$ is the

transferable signal of the training rule on the target task. In the Bernoulli agreement experiment, majority voting is the optimal polarity test, so this is the Bayes accuracy of the reference learner within $\mathcal{H}$.

At $K = 1$, $P_c(1, s) = s$, which yields $\Delta_{\mathrm{ICL}}(1, s) = 2(s - 1/2)^2$, so the one-shot gain is quadratic in the similarity margin. For $s > 1/2$, $K \to \infty$ gives $P_c \to 1$ by the law of large numbers, and thus $\Delta_{\mathrm{ICL}} \to s - 1/2$, corresponding to the asymptote in Figure 2. By Hoeffding's inequality, $1 - P_c(K, s) \leq \exp\left[-2K(s - 1/2)^2\right]$, meaning the context size required to drive the polarity error below δ scales as $K = \mathcal{O}\big(\log(1/\delta)/(s - 1/2)^2\big)$; lower similarity lowers the asymptote and slows the approach to it.

A trained network need not attain this accuracy. We write its gain as $\Delta_{\mathrm{model}} \approx \eta\, \Delta_{\mathrm{ref}}$ with an efficiency coefficient η fitted per learner and setting. Lemma 1 shows that errors which act as label-symmetric channels produce this form with $\eta \leq 1$ and with a constant η at large K. The fitted values are reported in Section 4.1 and Section 4.2.

Prediction 1 (the few-shot surface). For a single training rule, the gain at every support size follows the surface of Theorem 2 up to a multiplicative efficiency η. In particular, the one-shot gain is quadratic in the agreement with each evaluation target, and the gain at large K approaches $\eta\,(s - \frac{1}{2})$. Against the target-averaged axis of the structural ladder, the reference learner's one-shot exponent is 1.82 (App. D.1). Section 4.1 tests this prediction.

3.3 Selection and averaging over multiple training rules

The analysis so far assumes a single training rule. When the training pool contains T rules, $\mathcal{R} = \{r_1, \ldots, r_T\}$, the hypothesis class becomes $\mathcal{H} = \{\pm r_1, \ldots, \pm r_T\}$. Under the label-symmetric likelihood of Section 3.1 and a uniform prior, the posterior weight of a hypothesis h depends on the support set S only through its match count m_h, the number of support labels that h predicts correctly, $P(h \mid S) \propto \exp(\beta\, m_h)$ with $\beta = \log[(1-\varepsilon)/\varepsilon] > 0$ (Proposition 2). A multi-rule learner therefore weights the training rules and their polarities by their agreement with the support set.

Two regimes follow. For each rule r_j, let r_j^* denote the polarity aligned with the evaluation task e, and let $s_j = P[r_j^*(X) = e(X)]$ be its population agreement. If one rule r_{j^*} has a larger agreement than every other member, its expected advantage in match count grows linearly in K, whereas sampling fluctuations grow as $\sqrt{K}$. The posterior then concentrates on r_{j^*}, and the asymptotic accuracy equals $\max_j s_j$. We refer to this regime as *selection*. If the members have matched agreements, their match counts differ only by sampling fluctuations of order $\sqrt{K}\,\sigma$, with σ^2 the variance of the pairwise agreement difference. The aligned members then receive nearly equal posterior weight when $\beta\sqrt{K}\sigma \ll 1$, and the prediction is their node-wise majority vote. We refer to this regime as *averaging*. For a fixed β it occupies an intermediate range of K, since as $K \to \infty$ the posterior concentrates on the member with the largest empirical match count (Proposition 2(ii)); the support size at which two members become distinguishable is stated in Proposition 2(iii).

In the averaging regime, the gain depends on the joint error structure of the members rather than on their number. Define the agreement of the members' node-wise majority vote with the evaluation task as $s_{\mathrm{maj}} = P[\mathrm{Maj}_j\, r_j^*(X) = e(X)]$. At matched agreements, independent member errors make s_{maj} exceed s_j (the jury theorem (Boland, 1989), Proposition 2(ii)), so the averaged predictor is more accurate than any single member; if the members err on the same nodes, $s_{\mathrm{maj}} = s_j$ and adding rules does not increase the transferable signal.

The two regimes therefore give two predictions, $\max_j s_j$ under selection and s_{maj} under averaging, both computed from labeled examples before training. Together with Proposition 1, which states that the surface of Theorem 2 depends on the task pair and the graph prior only through the agreement statistics, this yields two predictions.

Prediction 2 (inert diversity axes). If the agreement statistics with the evaluation task are unchanged, the gain of the reference learner is unchanged. Adding graph families or training rules that leave the agreement statistics unchanged therefore brings no gain.

Prediction 3 (complementary diversity). In the averaging regime the gain of the reference learner tends to $s_{\mathrm{maj}} - 1/2$ (Proposition 2(ii)), so a pool with matched member agreements improves transfer

when $s_{\mathrm{maj}} > s_j$. The pre-registered finite-context forecast applies Theorem 2 to the aggregate rule with agreement s_{maj} at $K = 16$, a surrogate for the full posterior (App. B.4).

4 EXPERIMENTS

4.1 TESTING THE THREE PREDICTIONS

Prediction 1 (the few-shot surface). This test uses the single-rule runs of the twelve cells of Section 2.2. In each cell one model class is trained once on each of the 40 training rules of one ladder, and evaluation records the accuracy at $K \in \{1, 4, 8, 16\}$, so every rule yields one curve of the gain against K. We compare each curve with the surface of Theorem 2 in three forms, the asymptote $s - \frac{1}{2}$ applied at every K, the zero-parameter surface itself, and that surface times one fitted efficiency η. Each fit is scored by the R^2 of the prediction against the data, which is negative when the prediction is worse than the mean.

At K=16 the 40 points of the Model A structural ladder fit the asymptote at $R^2 = 0.801$, and one fitted efficiency gives $\eta = 0.917$ (Figure 3a). On the mixed ladder and on B/C1/C2 the fitted efficiency is 0.79 to 0.96. Over all four support sizes the asymptote reaches $R^2 = 0.58$ at best and is negative on the mixed ladders. The zero-parameter surface fits the eight structural and feature cells at $R^2 = 0.81$ to 0.92, and one fitted efficiency brings the four mixed cells to 0.89 to 0.94. Table 6 reports all three forms in all twelve cells.

At K=1 the surface is quadratic in $s - \frac{1}{2}$. The fitted exponent on the structural ladders is 1.64 to 2.21, against 2 for a single evaluation target and 1.82 for the reference learner on the ladder's target-averaged axis. Dividing the gain of every (rule, K) point by $s - \frac{1}{2}$ leaves the vote factor $2P_c(K, s) - 1$, and the quotients lie on the predicted identity with slope 0.78 to 1.00 (Figure 3b). App. D.1 and App. D.2 report the bootstrap intervals, the feature-ladder exponents, the K=0 tie offset of B/C1/C2, and the intercept of the unconstrained line of Figure 3a.

Prediction 2 (inert diversity axes). Two experiments keep the evaluation tasks fixed, vary one axis of the training prior, and measure the gain at K=16. The first holds four training rules fixed and raises the number of graph families M in the training prior from 1 to 8, on all four models (5 seeds). No model gains with M. The second raises the number of training rules T from 1 to 16 on Model A and Model B (5 seeds). Every member of both pools uses all three structural statistics that the evaluation rules use, so predicate coverage is saturated at T=1 and constant in T, and the two pools differ only in their similarity to those rules (App. D.3). The similar pool gives $\Delta_{\mathrm{ICL}} \approx 0.35$ and the dissimilar pool 0.06 to 0.11 at every T, and both curves are flat in T (Figure 3c). App. D.3 and App. D.4 report the slopes and correlations, and the grid that crosses polarity randomization with pool similarity.

Prediction 3 (complementary diversity). This experiment uses three training pools of T=3 rules whose members have matched similarity to the evaluation tasks ($s_j \approx 0.79$) and differ in error structure. The members of the complementary pool C err on different nodes, $s_{\mathrm{maj}} = 0.840$, and the members of the redundant pool R err on the same nodes, $s_{\mathrm{maj}} = 0.793$. A single-rule control S (T=1) provides the third arm. Model A is trained on each pool (5 seeds for C and R, 3 for S) and the gain is measured at K=16. Before training, selection predicts $\Delta_{\mathrm{ICL}} \approx 0.27$ to 0.28 for all three arms and averaging separates C at 0.319 (Table 8). The observed contrast $\Delta^C_{\mathrm{ICL}} - \Delta^R_{\mathrm{ICL}} = +0.049$ matches the averaging prediction of $+0.047$, against -0.006 under selection (Figure 3d). App. D.5 reports the training-budget controls.

4.2 TRANSFER ON REAL GRAPHS

We first score the law out of sample on the twelve cells of Section 2.2. A rule and its negation are replicates (Section 2.1), so the held-out unit is a weight direction rather than a single rule. In each cell the one-parameter form is fitted on the rules of the other directions and forecasts the gain of the held-out one. This gives $Q^2 = 0.76$ to 0.92 in all twelve cells, and the zero-parameter surface, with no fitted quantity, forecasts the Model A structural ladder at $Q^2 = 0.885$ (Table 9).

We then replace the synthetic ingredients of the testbed one at a time by their real counterparts, keeping disconnection and polarity randomization. Episode graphs become BFS subgraphs of five

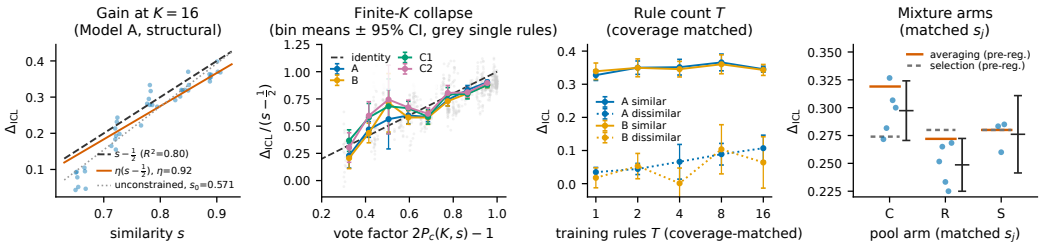

Figure 3: **The three predictions.** (a) The gain at K=16 on Model A structural runs, with the zero-parameter line ($R^2 = 0.80$), one-parameter form ($\eta = 0.92$), and the unconstrained line whose intercept is derived in App. D.2. (b) Finite-K collapse (Theorem 2), $\Delta_{\mathrm{ICL}}/(s - \frac{1}{2})$ against the vote factor for all rules and $K \in \{1, 4, 8, 16\}$, four models. (c) Rule count T at matched coverage, with the similar pool flat and high and the dissimilar pool flat and low. (d) The mixture experiment, per-seed Δ_{ICL} for the C, R, and S arms against the two predictions recorded before training.

real networks, Cora, CiteSeer, and PubMed (Yang et al., 2016) and the heterophilous Roman-empire and Amazon-ratings (Platonov et al., 2023), and similarity is re-measured on real subgraphs under the protocol of Section 2.1 (App. E.1). The three settings below use Model A.

Real structure and real features. The first two settings train a T=8 pool at each of five similarity levels on every dataset (3 seeds). The first uses synthetic features with structural rules, the second the dataset's real features with feature rules, and both correlate the gain at K=16 with the pool's mean member agreement, which is the registered default axis. The correlation between the gain and similarity is $+0.93$ to $+0.99$ on all five datasets with real features. With real structure it is $+0.98$ to $+0.99$ on the three homophilous datasets (Cora, CiteSeer, PubMed) and falls to $+0.73$ on Roman-empire and $+0.37$ on Amazon-ratings.

Both low values come from the choice of pool statistic. Section 3.3 summarizes a pool by s_{maj} under averaging and by $\max_j s_j$ under selection, and the real pools are in the averaging regime at K=16 (App. B.4). On the heterophilous graphs the members err on different nodes more often, so the majority vote agrees with the evaluation task more than the mean member does, and the mean no longer orders the five levels. On the s_{maj} axis the two heterophilous datasets rise to $+0.86$ and $+0.90$ and the three homophilous ones stay at $+0.85$ to $+0.98$, so the structural law holds on all five datasets under the pool statistic that the averaging regime prescribes. The selection statistic $\max_j s_j$ gives $+0.24$ on Amazon-ratings. Per-dataset values under the three axes are in Table 11.

Scored as a forecast, η fitted only on the synthetic feature ladder predicts all 25 real points (five datasets at five levels) at $Q^2 = 0.872$ with no real-graph fitting (Figure 4a). The leave-one-dataset-out forecasts of both settings are in Table 9.

Real labels. The third setting uses the dataset's own node classes as tasks. On the 18 classes of Roman-empire a task labels a node by membership in a 9-class subset. The reference task is one fixed subset, and the training tasks are 40 subsets with graded overlap with the reference. Similarity is the polarity-symmetrized label agreement of Section 2.1, balanced with respect to the reference task to remove class imbalance and measured on real-feature subgraphs. Each training task trains one Model A (T=1, 3 seeds), evaluated on the reference task at K=16. The law holds at $r = +0.842$ with leave-one-task-out $Q^2 = 0.661$ (Figure 4b). On real labels s therefore forecasts the ranking of tasks with no training. The absolute gain still requires one efficiency per learner and setting, which is $\eta = 0.45$ when fitted here and does not transfer from the synthetic or real-feature settings ($Q^2 < 0$ when imported). Part of the decrease is attributable to the axis (App. E.5).

Identifiability. The law presupposes a rule that the demonstration graphs and the query graph share. Synthetic rules satisfy this by construction, and a dataset's own class labels need not. Before training we therefore fit a linear probe to the reference task on each real-feature subgraph. The invariant mass $\hat{\lambda}^2$ is the consistency of these probe directions across subgraphs (App. E.5). When the directions vary from subgraph to subgraph, an in-context learner has no shared rule to recover, and its training loss stays at $\ln 2$ (Proposition 4). From the measured values we pre-registered no transfer on Amazon-ratings ($\hat{\lambda}^2 = 0.06$) and transfer on CiteSeer (0.53), and both were confirmed, Amazon-ratings flat

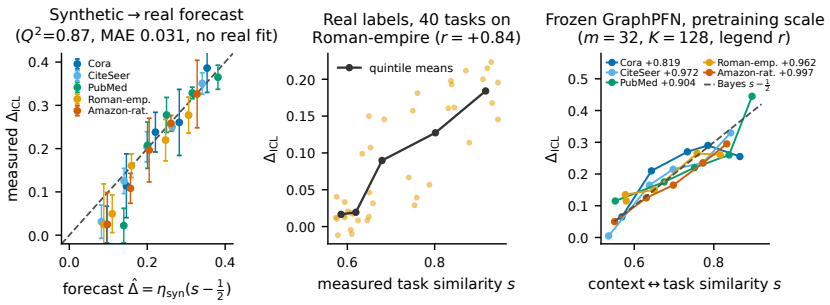

Figure 4: **Transfer on real graphs.** (a) Synthetic-to-real forecast. η from the synthetic feature ladder predicts all 25 real (dataset $\times$ level) points at $Q^2 = 0.872$. (b) Real labels, 40 node-classification tasks on Roman-empire ($r = +0.842$). (c) The frozen GraphPFN checkpoint at its pretraining scale (m=32, K=128), Δ_{ICL} against context–task similarity for feature rules on the real features of five datasets, with the Bayes line $s - \frac{1}{2}$ and per-dataset r in the legend.

with training loss at $\ln 2$ and CiteSeer at $r = +0.628$. Cora (0.085) was predicted not to transfer and measures $r = +0.821$. The law therefore also holds where $\hat{\lambda}^2$ is small, because $\hat{\lambda}^2$ bounds only the component that is invariant across graphs, and Cora transfers through a component that depends on graph structure. Two further quantities are measurable before training, s_{maj} for multi-rule pools and $\hat{\lambda}^2$ for the graph-invariant component of a real-label task (Table 12).

4.3 External validation on a frozen foundation model

The last test trains no model. We run the public GraphPFN checkpoint (Eremeev et al., 2025), about 20M parameters, pretrained on synthetic attributed graphs and frozen, on our episodes. GraphPFN combines a tabular transformer, whose attention over samples spans all nodes of the input, with graph adapters that pass messages along edges. We feed the support and query graphs as a disconnected union. The union has no edge between a demonstration graph and the query graph, so the adapters cannot propagate labels and the support labels reach the query only through the sample attention.

The pretraining pool of the frozen checkpoint is fixed, so the similarity axis is placed in the context instead. The support nodes are relabeled with a graded rule r of the feature ladder while the query keeps its evaluation-rule label e. The node input is the eight leading principal components of the dataset's node features, with no structural descriptor (App. F.1). A learner that recovers the context rule and its polarity from the support attains accuracy approaching $s(r, e)$.

Episodes are built at the model's pretraining scale, m=32 demonstration graphs, unions of about 1,600 nodes, and K=128, with five similarity levels and 200 episodes per level on each of the five datasets. The gain is measured against the support-permutation control on the same episodes, which is at chance (0.499 pooled over the 25 dataset $\times$ level points). The frozen model reproduces the law on all five datasets, $r = +0.82$ to $+0.997$ (Figure 4c), with efficiencies $\Delta_{\mathrm{ICL}}/(s - \frac{1}{2})$ of 0.88 to 0.97. At the default episode size (m=4, K=16) the law holds on four of the five datasets and the same model gives $\eta = 0.25$ to 0.54. The efficiency rises with the number of labeled examples rather than with the size of the union (App. F.2).

5 Conclusion

This paper identifies a general law for in-context learning on graphs. The in-context gain on a new task is governed by its polarity-symmetrized label agreement with the training task, and a Bayesian analysis predicts the whole few-shot curve from this agreement with no fitted parameter. The law holds for four model classes, five real graph datasets, and a frozen graph foundation model that we did not train. Because the agreement is measured from labeled examples alone, a source task can be scored before any training run. Future work could use it to choose the source tasks of a pretraining corpus for a graph foundation model.

AI USE STATEMENT

Large language model assistants were used to polish the text and to write the figure-generation and data-analysis scripts. The authors reviewed all generated material and take responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

The anonymized code release contains the testbed, the four model classes, the analysis scripts, the Hydra configurations, and the fixed seeds. App. G gives the architectures, the training budget, and the episode parameters. The real-graph protocol is in App. E.1 and the protocol for the frozen GraphPFN checkpoint is in App. F.1.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6430–6439, 2019.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Aryaman Arora, Dan Jurafsky, Christopher Potts, and Noah D Goodman. Bayesian scaling laws for in-context learning. *arXiv preprint arXiv:2410.16531*, 2024.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999.

Philip J Boland. Majority systems and the condorcet jury theorem. *Journal of the Royal Statistical Society Series D: The Statistician*, 38(3):181–189, 1989.

William Brown. Some experimental results in the correlation of mental abilities 1. *British Journal of Psychology, 1904-1920*, 3(3):296–322, 1910.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891, 2022.

Paul Erdős and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

Dmitry Eremeev, Oleg Platonov, Gleb Bazhenov, Artem Babenko, and Liudmila Prokhorenkova. Graphpfn: A prior-data fitted graph foundation model. *arXiv preprint arXiv:2509.21489*, 2025.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in neural information processing systems*, 35:30583–30598, 2022.

Chase Goddard, Lindsay M Smith, Vudtiwat Ngampruetikorn, and David J Schwab. When can in-context learning generalize out of task distribution? *arXiv preprint arXiv:2506.05574*, 2025.

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

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Kržmanc, Daniel Zeng, Percy S Liang, and Jure Leskovec. Prodigy: Enabling in-context learning over graphs. *Advances in Neural Information Processing Systems*, 36:16302–16317, 2023.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *International Conference on Learning Representations*, volume 2024, pp. 3209–3280, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Mary Letey, Jacob Zavatone-Veth, Yue Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 128723–128761, 2026.

Ziqian Lin and Kangwook Lee. Dual operating modes of in-context learning. *arXiv preprint arXiv:2402.18819*, 2024.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International conference on learning representations*, volume 2024, pp. 20188–20210, 2024.

Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S Yu, et al. Graph foundation models: Concepts, opportunities and challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):5023–5044, 2025.

Yue M Lu, Mary Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International conference on machine learning*, pp. 7294–7305. PMLR, 2020.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 66381–66433, 2025.

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

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in neural information processing systems*, 36:14228–14246, 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

William Fleetwood Sheppard. On the application of the theory of error to cases of normal distribution and normal correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 192:101–531, 1899.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in neural information processing systems*, 36:27801–27819, 2023.

Aaditya K Singh, Ted Moskovitz, Sara Dragutinovic, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. Strategy coopetition explains the emergence and transience of in-context learning. *arXiv preprint arXiv:2503.05631*, 2025.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Charles Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Qixun Wang, Yifei Wang, Xianghua Ying, and Yisen Wang. Can in-context learning really generalize to out-of-distribution tasks? In *International Conference on Learning Representations*, volume 2025, pp. 83553–83574, 2025.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393 (6684):440–442, 1998.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, volume 2024, pp. 56587–56641, 2024.

Daniel Wurgaft, Ekdeep S Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah Goodman. In-context learning strategies emerge rationally. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 118425–118485. Curran Associates, Inc., 2025. doi: 10.52202/085713-3950. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/abb22b1e4b4ab2988862be9d7ed083f9-Paper-Conference.pdf.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

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

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International conference on machine learning*, pp. 12133–12143. Pmlr, 2021.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Jianan Zhao, Zhaocheng Zhu, Mikhail Galkin, Hesham Mostafa, Michael Bronstein, and Jian Tang. Fully-inductive node classification on arbitrary graphs. In *International Conference on Learning Representations*, volume 2025, pp. 23968–23986, 2025.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

A Related work

Determinants of in-context learning. Controlled studies train transformers from scratch on parametric task families (Garg et al., 2022; Akyürek et al., 2022; Von Oswald et al., 2023) and attribute the emergence of ICL to properties of the pretraining distribution, namely burstiness and a large number of rarely occurring classes (Chan et al., 2022) and the number of distinct pretraining tasks (Raventós et al., 2023; Wu et al., 2024). Later analyses qualify this account. Letey et al. (2026) show in a solvable linear-attention model (Lu et al., 2025) that ICL generalization is governed by the alignment between the pretraining and test task distributions, and that added pretraining diversity helps or hurts depending on that alignment. This is the closest antecedent of our law. Their theory is derived for linear regression, and their experiments show that the same measure predicts the error of softmax-attention transformers trained on the same tasks. For Gaussian linear threshold rules, Proposition 3 relates our agreement axis to the cosine between rule weight vectors, which is the single-rule analogue of their covariance-level alignment measure. Goddard et al. (2025) map a phase diagram over task diversity and task count for in-context linear regression, find a sharp transition from a specialized to a generalizing solution, and observe the same transition on nonlinear regression, whereas we observe no transition on continuous metrics along our axis. Wang et al. (2025) argue that on an out-of-distribution task a model selects the pretraining function with the lowest error on the test context. Training-dynamics studies describe when ICL emerges, persists, or is displaced by competing strategies (Singh et al., 2023; 2025; Park et al., 2025; Wurgaft et al., 2025). Our analysis extends this account to graphs through label agreement measured from task examples, a parameter-free finite-context law, and forecast evaluation on real graphs and a frozen external model. Whether a model reads the in-context labels at all is tested in language models with flipped or random labels (Wei et al., 2023; Kossen et al., 2024). In our testbed polarity randomization makes label reading necessary by construction.

Emergence and its metrics. Reported emergent abilities (Wei et al., 2022) can be artifacts of discontinuous metrics (Schaeffer et al., 2023). The main result of this paper is a smooth positive law. The law is replicated on continuous metrics (App. C.1), and every statement about the absence of emergence is restricted to this testbed. When the in-weights channel is reopened, ICL can be emergent (Singh et al., 2025).

Transferability estimation. A separate literature predicts fine-tuning transfer. LEEP (Nguyen et al., 2020) and LogME (You et al., 2021) pass the target data through the source checkpoint, task embeddings (Achille et al., 2019) pass it through a fixed probe network, and optimal-transport dataset distances (Alvarez-Melis & Fusi, 2020) compare labeled datasets without a model. We predict in-context gain from labeled examples of the source and target tasks before the learner is trained. On our runs (Table 3), LogME is the strongest alternative, trails s in every cell, and requires the checkpoint. LEEP is constant on B/C1/C2 because their tied zero-shot logits give a uniform predictive distribution (Table 3).

Graph in-context learning. PRODIGY performs few-shot node classification through prompt-graph readouts (Huang et al., 2023), and graph prompting unifies tasks through prompt graphs (Liu et al., 2024). Prior-data fitted networks (Müller et al., 2021) bring the meta-ICL recipe to tabular and graph data (Hollmann et al., 2022; 2025; Eremeev et al., 2025). Graph foundation models aim at transferable node-level capability (Liu et al., 2025; Zhao et al., 2025). These works report accuracy in settings where support and query nodes come from the same graph with their neighborhoods, so propagation, homophily, and feature shortcuts are available and the measured capability is not necessarily in-context inference. Our testbed separates in-context inference from these shortcuts, Model B and Model C1/Model C2 bring the corresponding readout families into the testbed, and Section 4.3 evaluates the released GraphPFN checkpoint under the same guarantees.

Bayesian accounts. ICL has been formalized as implicit Bayesian inference of a latent concept (Xie et al., 2021), with Bayesian scaling laws along the shot axis (Arora et al., 2024) and a distinction between task retrieval and task learning (Lin & Lee, 2024). Our law is stated on the orthogonal train–test similarity axis and is resolved in K as well.

B THEORETICAL DETAILS

Section 3 states two theorems and three predictions. This appendix proves the theorems (App. B.1), proves the isolation guarantee used in Section 2 (App. B.2), and adds three results that the main text uses without proof. App. B.3 states when the efficiency coefficient η is exact and shows that the surface depends on the task pair only through the agreement statistics. App. B.4 proves the selection and averaging regimes of multi-rule pools and evaluates the regime criterion on every pool used in this paper. App. B.5 relates the agreement axis to weight-space alignment.

B.1 PROOFS FOR SECTION 3

Write $X = (G, v)$ for a node together with its graph, and encode binary labels as $\{-1, +1\}$. An episode presents $Y = \varsigma e(X)$ at evaluation and $Y = \varsigma r(X)$ during single-rule training, where $\varsigma \in \{-1, +1\}$ is uniform and independent of the graphs. The support sampling rule is invariant under a simultaneous exchange of the two class names. In the protocol of Section 2.1, this invariance follows by exchanging the uniformly random starting class of the round-robin sampler. Consequently, unlabeled graphs and support identities leave the polarity prior uniform.

For the finite-context calculation, fix one evaluation task and choose a global orientation r^* with agreement $s = P[r^*(X) = e(X)] \geq \frac{1}{2}$. Assume that the support indicators $A_i = \mathbf{1}\{r^*(X_i) = e(X_i)\}$ are independent $\mathrm{Bernoulli}(s)$ variables and that the query indicator A_q is an independent variable with the same law. These assumptions specify the Bernoulli reference experiment used below. The within-graph sampling extension is given in Remark 3.

Proof of Theorem 1. Let Z collect the unlabeled episode data, including the query and support identities, and let S contain the observed support labels. Put $p_y = P(Y_q = y \mid Z, S)$. Conditional cross-entropy decomposes as

$$-\sum_y p_y \log q_y = -\sum_y p_y \log p_y + \sum_y p_y \log \frac{p_y}{q_y} = H(p) + \mathrm{KL}(p\|q).$$

Taking expectations and using nonnegativity of KL divergence shows that the population minimum is attained exactly when $q = p$ almost surely under the training distribution.

For a fixed training rule r, the conditional label law is determined by the episode sign, so

$$p_y = \sum_{a \in \{-1,+1\}} P(\varsigma = a \mid Z, S)\mathbf{1}\{y = ar(X_q)\}.$$

At $K = 0$, polarity invariance of Z gives $P(\varsigma = a \mid Z) = 1/2$, hence $p_{+1} = p_{-1} = 1/2$. At $K \geq 1$, each training support pair satisfies $Y_i = \varsigma r(X_i)$, so $\varsigma = Y_i r(X_i)$ is determined by one pair. All remaining pairs give the same sign. The posterior is therefore a point mass and the population predictive assigns probability one to $\varsigma r(X_q)$. □

Proof of the majority-vote equivalence (Section 3.1). The reference learner scores held-out support labels by an independent binary symmetric channel with error probability $\varepsilon \in (0, \frac{1}{2})$. For a hypothesis h, let $m_h = \sum_{i=1}^K \mathbf{1}\{h(X_i) = Y_i\}$. Its likelihood is

$$L(h; S) = (1 - \varepsilon)^{m_h} \varepsilon^{K - m_h}.$$

The hypotheses $h^+ = r$ and $h^- = \bar{r} = -r$ (under the ± 1 encoding) have equal priors and complementary match counts m and $K - m$. Their posterior log-odds are

$$\log \frac{P(h^+ \mid S)}{P(h^- \mid S)} = (2m - K) \log \frac{1 - \varepsilon}{\varepsilon}.$$

The logarithmic factor is positive. The decision selects h^+ for $m > K/2$, selects h^- for $m < K/2$, and randomizes uniformly at equality. Since the two hypotheses predict opposite labels at every query, this is also the classification decision of their posterior predictive. The classification decision is independent of ε throughout the stated interval. Any common strictly increasing transformation of the match counts gives the same ordering. □

Proof of Theorem 2. Let $M_K = \sum_{i=1}^{K} A_i$. Independence gives $M_K \sim \mathrm{Binomial}(K, s)$, and the probability of selecting the aligned polarity is

$$P_c(K, s) = \sum_{m>K/2} \binom{K}{m} s^m (1-s)^{K-m} + \frac{1}{2}\mathbf{1}\{K \text{ is even}\}\binom{K}{K/2} s^{K/2}(1-s)^{K/2}.$$

Write $C = 1$ when the aligned polarity is selected. The correctness indicator is $CA_q + (1-C)(1-A_q)$. Since C depends on support votes and independent tie randomness, it is independent of A_q. Therefore

$$\begin{aligned}\mathrm{Acc}(K) &= P_c s + (1-P_c)(1-s) \\ &= \frac{1}{2} + (2P_c - 1)\left(s - \frac{1}{2}\right).\end{aligned}$$

Subtracting $\mathrm{Acc}(0) = 1/2$ proves the formula, including $K = 0$ under the convention $P_c(0, s) = 1/2$.

For optimality in the Bernoulli reference experiment, write $B_i = \mathbf{1}\{r(X_i) = Y_i\}$ and let $\Theta = +$ mean that r has the aligned episode polarity. Conditional on $\Theta = +$ the B_i are independent $\mathrm{Bernoulli}(s)$ variables, and conditional on $\Theta = -$ they are independent $\mathrm{Bernoulli}(1-s)$ variables. The two signs have equal priors. For $1/2 < s < 1$, their likelihood ratio is

$$\frac{P(B_1, \ldots, B_K \mid \Theta = +)}{P(B_1, \ldots, B_K \mid \Theta = -)} = \left(\frac{s}{1-s}\right)^{2\sum_i B_i - K}.$$

At each observed vector, selecting the sign with larger posterior probability maximizes conditional probability of correct polarity recovery. This gives the majority rule. For any polarity rule based on these observations, the query accuracy is $1/2 + (2P[\widehat{\Theta} = \Theta] - 1)(s - 1/2)$, so the same decision maximizes query accuracy. At $s = 1/2$ both signs give accuracy $1/2$; at $s = 1$, one support observation recovers the sign exactly. □

Proof of the limits of the surface (Section 3.2). For $K = 1$, the aligned polarity is chosen exactly when $A_1 = 1$, giving $P_c(1, s) = s$ and $\Delta_{\mathrm{ICL}}(1, s) = 2(s - 1/2)^2$. For $s > 1/2$, the law of large numbers gives $M_K/K \to s$ almost surely, so $P_c(K, s) \to 1$ and $\Delta_{\mathrm{ICL}}(K, s) \to s - 1/2$. At $s = 1/2$, binomial symmetry gives $P_c(K, 1/2) = 1/2$ for every K, and the gain is zero.

Put $d = s - 1/2 > 0$. Hoeffding's inequality gives

$$1 - P_c(K, s) \le P\left[\frac{M_K}{K} - s \le -d\right] \le e^{-2Kd^2}.$$

Thus a polarity error at most δ is ensured by $K \ge \log(1/\delta)/(2d^2)$. The vote factor and gain satisfy

$$1 - (2P_c - 1) \le 2e^{-2Kd^2}, \qquad 0 \le d - \Delta_{\mathrm{ICL}}(K, s) \le 2de^{-2Kd^2}.$$

In particular, the vote factor is within δ of one when $K \ge \log(2/\delta)/(2d^2)$. □

B.2 The isolation guarantee

Corollary 1 establishes the expected accuracy of predictors using label-free episode data. Remark 1 gives the corresponding finite-sample accuracy when a model always returns one class at $K = 0$.

Corollary 1 (chance is exact under polarity randomization)**.** *Let π be a uniform random permutation of C classes, independent of the unpermuted episode, and assume that query and support sampling is invariant under class relabeling. Every predictor measurable with respect to the label-free episode data, with independent randomization allowed, has expected accuracy $1/C$ at every K. In particular, $\mathrm{Acc}(0) = 1/C$. Binary polarity randomization is the case $C = 2$.*

Proof of Corollary 1. Let Z contain all graphs and features, the query and support identities, and K. Condition additionally on the unpermuted task e. Sampling invariance implies $P(\pi = \tau \mid Z, e) = 1/C!$ for every permutation $\tau \in S_C$. For a fixed base label $b = e(X_q)$ and a presented class c, exactly $(C-1)!$ permutations satisfy $\tau(b) = c$. Hence

$$P(Y_q = c \mid Z, e) = \frac{1}{C!} \sum_{\tau \in S_C} \mathbf{1}\{\tau(b) = c\} = \frac{1}{C}.$$

Averaging over e preserves this conditional probability. If R is independent prediction randomness and $\widehat{Y} = f(Z, R)$, then $P(\widehat{Y} = Y_q \mid Z, R) = 1/C$, and taking expectations proves the claim. For the binary protocol, exchanging the random starting class preserves the support-identity distribution, and selecting the query class uniformly preserves the query-identity distribution. These observations verify the sampling assumption for every K.

Predictions using pretrained weights, query features, or unlabeled query structure are functions of Z. Label propagation is excluded by the graph construction, because every path from a support node remains inside a demonstration component. □

Remark 1 (the $K{=}0$ tie offset on B/C1/C2). Corollary 1 is a statement in expectation over the episode process. On B/C1/C2 the $K{=}0$ logits tie exactly and the argmax returns a constant class, so the predictor scores the realized label marginal of the fixed 600-episode evaluation sample, and this value is what the evaluation records. Every B/C1/C2 run scores $321/600 = 0.535$ on the structural stream and $320/600 \approx 0.533$ on the feature and mixed streams, across the three architectures and all forty rules. The shared evaluation sample determines this constant accuracy. The metric readout of Model A does not tie, and its $\mathrm{Acc}(0)$ scatters around $\frac{1}{2}$ per run. Redrawing the evaluation seed re-centers the offset at chance. Figure 2 subtracts the measured zero-shot accuracy of each run. The comparisons with the theoretical surface (App. D.1) subtract $\frac{1}{2}$ on B/C1/C2, which shifts their gains by the tie offset and leaves every correlation unchanged.

B.3 The efficiency coefficient

Section 3.2 models a realized learner as the Bayes surface multiplied by an efficiency η. Lemma 1 gives the exact finite-context efficiency and its large-context limit. Proposition 1 states that the surface depends on the task pair and on the graph prior only through the agreement statistics, which is the premise of Predictions 2 and 3.

Lemma 1 (Composition of efficiency factors)**.** *Consider the majority reference learner with a rule estimate $\hat{r}$. At each support and query node, the estimate agrees with the aligned rule r^* with probability $q \in [1/2, 1]$. Assume these agreement events are independent across nodes and independent of the agreements between r^* and e. Independently, flip the learner's final answer with probability $\gamma \in [0, 1/2]$. Define $\tilde{s} = qs + (1-q)(1-s)$ and $V_K(s) = 2P_c(K, s) - 1$. Then*

$$\Delta_{\mathrm{ICL}}(K) = (1-2\gamma)V_K(\tilde{s})(\tilde{s} - \tfrac{1}{2}), \qquad \tilde{s} - \tfrac{1}{2} = (2q-1)(s - \tfrac{1}{2}).$$

The asymptotic efficiency is $\eta_\infty = (2q-1)(1-2\gamma)$. For $K \geq 1$ and $s > 1/2$, the exact finite-context efficiency relative to Theorem 2 is

$$\eta_K = \eta_\infty \frac{V_K(\tilde{s})}{V_K(s)}, \qquad 0 \leq \eta_K \leq \eta_\infty \leq 1.$$

At $K = 1$, $\eta_1 = (2q-1)^2(1-2\gamma)$.

Proof. The estimated rule is correct when both $\hat{r} = r^*$ and $r^* = e$, or when both equalities fail. Independence gives $P(\hat{r} = e) = qs + (1-q)(1-s) = \tilde{s}$. The assumptions also preserve independence between the estimated rule's support and query agreements. Applying Theorem 2 with $\tilde{s}$ gives the gain before the output flip.

If the pre-flip accuracy is a, the post-flip accuracy is

$$(1-\gamma)a + \gamma(1-a) = \frac{1}{2} + (1-2\gamma)(a - \tfrac{1}{2}).$$

The zero-shot accuracy remains $1/2$, so the output channel multiplies the gain by $1-2\gamma$. Substitution proves the first display. Dividing by $V_K(s)(s - 1/2)$ gives η_K. Since $1/2 \leq \tilde{s} \leq s$ and V_K is

nonnegative and increasing on $[1/2, 1]$, the stated bounds follow. At $K = 1$, $V_1(s) = 2s - 1$, giving the additional factor $2q - 1$.

For $\tilde{s} > 1/2$, both vote factors converge to one, so $\eta_K \to \eta_\infty$. More explicitly, the constant-efficiency approximation exceeds the exact gain by

$$D_K = (1 - 2\gamma)(\tilde{s} - \tfrac{1}{2})\big[V_K(s) - V_K(\tilde{s})\big],$$

with $0 \le D_K \le 2(1 - 2\gamma)(\tilde{s} - 1/2)e^{-2K(\tilde{s}-1/2)^2}$. If $q = 1/2$, $\gamma = 1/2$, or $s = 1/2$, the exact gain and its asymptotic expression are both zero. □

Proposition 1 (Sufficiency and identification by agreement)**.** *Fix the channel parameter and prior of the reference learner. The joint law of its posterior mass on the aligned polarity, predictive probability of the correct query class, and correctness indicator depends on the task pair and graph distribution through the joint law of $(A_1, \dots, A_K, A_q)$. Under the independent Bernoulli assumptions, s determines this law. For a pool, the corresponding sufficient object is the joint array of all members' agreements at all support and query nodes. For each fixed $K \ge 1$, the gain in Theorem 2 is strictly increasing on $s \in [1/2, 1]$.*

Proof. Set $M = \sum_i A_i$ and $\beta = \log((1-\varepsilon)/\varepsilon)$. The posterior probability of the aligned hypothesis is

$$p_+(M) = \frac{1}{1 + \exp[-\beta(2M - K)]}.$$

Its predictive probability of the true query class is $A_q p_+(M) + (1 - A_q)(1 - p_+(M))$. The classification correctness is $CA_q + (1 - C)(1 - A_q)$, with C the majority decision and independent randomized ties. Thus all three quantities are functions of the agreement array and the fixed channel. For a pool, each weight is a function of one member's match count and each query contribution is determined by that member's query agreement. The same argument applies to the joint array.

To establish strict monotonicity, write $K = 2n + 1$ when K is odd. Differentiating the binomial tail and cancelling consecutive terms gives

$$\frac{\partial P_c(2n+1, s)}{\partial s} = (2n+1)\binom{2n}{n} s^n (1-s)^n > 0, \qquad 0 < s < 1.$$

For $K = 2n$, conditioning on the last vote gives $P_c(2n, s) = P_c(2n-1, s)$. Indeed, the contributions from a preceding count of n and $n - 1$ cancel, since $(1 - s)P[\mathrm{Bin}(2n - 1, s) = n] = sP[\mathrm{Bin}(2n - 1, s) = n - 1]$. Consequently $V_K(s) > 0$ and $V_K'(s) > 0$ for $1/2 < s < 1$, and

$$\frac{d}{ds}\big[V_K(s)(s - \tfrac{1}{2})\big] = V_K(s) + (s - \tfrac{1}{2})V_K'(s) > 0.$$

The gain is zero at $s = 1/2$ and continuous at both endpoints, proving strict increase on the closed interval. □

B.4 Multiple training rules

Section 3.3 states the selection and averaging regimes of a multi-rule pool. Proposition 2 proves them and gives the criterion that separates them, Remark 2 evaluates the criterion on every pool used in this paper, and a further remark states the effect of within-graph correlation.

Proposition 2 (Selection and averaging regimes)**.** *Fix a finite pool of $T \ge 2$ binary rules. Choose their global orientations r_j^* so that $s_j = P[r_j^*(X) = e(X)] \ge 1/2$. Assume the agreement vectors $A_i = (A_{i1}, \dots, A_{iT})$ are independent and identically distributed across support nodes, and that the query vector A_q is an independent draw from the same law. Give the $2T$ signed hypotheses equal prior mass and use likelihood weights $\exp(\beta_K m_h)$ with $\beta_K > 0$.*

(i) *Suppose j^* uniquely maximizes s_j, $s_{j^*} > 1/2$, and put $\delta = \min_{j \ne j^*}(s_{j^*} - s_j) > 0$. If $\sqrt{K}\,\delta \to \infty$ and $\beta_K K\delta \to \infty$, the posterior mass on the aligned j^* tends to one and classification accuracy tends to s_{j^*}. Both conditions hold for fixed $\beta_K = \beta > 0$ and a fixed positive gap.*

(ii) *Suppose $s_j = s \in (1/2, 1)$ for every j, and define $\sigma_{jj'}^2 = \mathrm{Var}(A_j - A_{j'})$ and $\sigma_{\max} = \max_{j,j'} \sigma_{jj'}$. If every distinct pair has $\sigma_{jj'} > 0$, $\beta_K\sqrt{K} \to \infty$, and $\beta_K K(2s - 1) \to \infty$, the*

posterior concentrates on the member with the largest empirical match count and accuracy tends to s. If

$$\beta_K\sqrt{K}\,\sigma_{\max} \to 0, \qquad \beta_K K(2s-1) \to \infty,$$

the posterior masses on the aligned members converge to $1/T$. The equal-weight classifier has accuracy

$$s_{\mathrm{maj}} = P\left[\sum_j A_{qj} > T/2\right] + \frac{1}{2}P\left[\sum_j A_{qj} = T/2\right].$$

The posterior classifier's accuracy converges to s_{maj} when T is odd, when query votes have zero tie probability, or when the joint member law is exchangeable. For independent member errors, $s_{\mathrm{maj}} = P_c(T, s) > s$ for every $T \geq 3$. Coincident error sets give $s_{\mathrm{maj}} = s$.

(iii) *For $\sigma_{jj'} > 0$, the empirical comparison of two members has standardized separation*

$$z_{jj'} = \frac{\sqrt{K}(s_j - s_{j'})}{\sigma_{jj'}}.$$

Under the central limit approximation, its probability of selecting j, with randomized ties, is $\Phi(z_{jj'})$. Reliable selection of the population ordering requires large absolute separation in this comparison. The parameter β_K controls posterior weight ratios while preserving the ordering of match counts.

Proof. The symmetric likelihood of a hypothesis with match count m_h is $\varepsilon^K \exp(\beta_K m_h)$, where $\beta_K = \log((1-\varepsilon)/\varepsilon)$. The common factor cancels from the posterior. Orient the episode polarity to $+1$ for the analysis, and put $m_j = \sum_i A_{ij}$. The aligned and opposite posterior weights are

$$w_{j,+} = \frac{e^{\beta_K m_j}}{Z_K}, \qquad w_{j,-} = \frac{e^{\beta_K (K-m_j)}}{Z_K}, \qquad Z_K = \sum_j \left(e^{\beta_K m_j} + e^{\beta_K (K-m_j)}\right).$$

The predictive probability of a correct query label is $\sum_j [w_{j,+}A_{qj} + w_{j,-}(1-A_{qj})]$. Thus the exact classification accuracy can be written as

$$\mathrm{Acc}(K) = \mathbb{E}\,\psi\left(\sum_j \left[e^{\beta_K m_j} - e^{\beta_K (K-m_j)}\right](2A_{qj}-1)\right), \qquad \psi(t) = \mathbf{1}\{t>0\} + \tfrac{1}{2}\mathbf{1}\{t=0\}.$$

This expression includes both polarities and all correlations between members at a node.

For *(i)*, let h^* denote the best aligned hypothesis, let a_h be the population agreement of any rival signed hypothesis, and set $d_h = s_{j^*} - a_h$. Every d_h is positive. Each difference in match counts is a sum of independent variables in $[-1, 1]$ with mean d_h. Hoeffding's inequality and a union bound give

$$P\left[\exists h \neq h^* : m_{h^*} - m_h < \frac{Kd_h}{2}\right] \leq \sum_{h \neq h^*} e^{-Kd_h^2/8}.$$

On the complementary event,

$$1 - w_{h^*} \leq \sum_{h \neq h^*} e^{-\beta_K K d_h/2}.$$

The minimum signed gap is at least δ, since an opposite hypothesis has agreement at most $1/2$. Both sums converge to zero under the stated conditions. Whenever $w_{h^*} > 1/2$, the posterior classifier agrees with h^* on every query. Boundedness of accuracy then gives the limit s_{j^*}.

For *(ii)*, matched means imply $m_j - m_{j'} = O_P(\sqrt{K}\,\sigma_{jj'})$, while $2m_j - K = K(2s-1) + O_P(\sqrt{K})$. Thus opposite weights vanish under $\beta_K K(2s-1) \to \infty$. For hard weighting, the multivariate central limit theorem gives a Gaussian limit for $((m_j - Ks)/\sqrt{K})_{j=1}^T$. Every pairwise difference has positive variance, so the limiting maximum is unique almost surely. The gap between the largest and second-largest counts, multiplied by β_K, therefore diverges in probability when $\beta_K\sqrt{K} \to \infty$. The largest aligned weight tends to one. Its index depends on support observations and is independent of A_q, and every possible selected member has query accuracy s.

For soft weighting, $\beta_K(m_j - m_{j'}) \to 0$ in probability for every pair. Hence the aligned weight ratios tend to one and each aligned weight tends to $1/T$. At a query with a strict majority, convergence of the predictive probability implies convergence of the class decision. Odd T always gives a strict majority. For even T with an exchangeable member law, a query tie divides the members into two groups of equal size. Exchanging these groups preserves the distribution of the support weights and reverses the vote margin. The conditional correctness on such a tie is therefore $1/2$, including randomized exact ties.

With independent member errors, the number of correct query votes is $\mathrm{Binomial}(T, s)$. The odd-count identity

$$P_c(2n+3, s) - P_c(2n+1, s) = \binom{2n+1}{n} s^{n+1}(1-s)^{n+1}(2s-1) > 0$$

follows by conditioning on the first $2n+1$ votes and the final two votes. Together with $P_c(1, s) = s$ and $P_c(2n, s) = P_c(2n-1, s)$, it proves strict improvement for all $T \geq 3$. If the member indicators coincide almost surely, their majority equals any member indicator.

For *(iii)*, put $D_i = A_{ij} - A_{ij'}$. Its mean is $\delta_{jj'} = s_j - s_{j'}$ and its variance is $\sigma^2_{jj'}$. In particular,

$$\sigma^2_{jj'} = s_j(1-s_j) + s_{j'}(1-s_{j'}) - 2\mathrm{Cov}(A_j, A_{j'}).$$

The central limit theorem applied to $\sum_i D_i$ gives the stated normal approximation. An exact one-sided bound for $\delta_{jj'} > 0$ is

$$P[m_j \leq m_{j'}] \leq \frac{K\sigma^2_{jj'}}{K\sigma^2_{jj'} + K^2\delta^2_{jj'}} = \frac{1}{1 + z^2_{jj'}},$$

by Cantelli's inequality. Finally, $\log(w_{j,+}/w_{j',+}) = \beta_K(m_j - m_{j'})$, so positive β_K preserves the empirical ordering. If $\sigma_{jj'} = 0$, the match-count difference is deterministic and can be evaluated directly. □

Finite-context forecast for an averaged rule. Let $b(X) = \mathrm{Maj}_j\, r^*_j(X)$ denote the equal-weight aggregate, with independent symmetric tie resolution, and let its agreement be s_{maj}. Applying the binary reference learner to the pair $\{b, -b\}$ with independent support and query agreements gives

$$\Delta_{\mathrm{aggregate}}(K) = \big(2P_c(K, s_{\mathrm{maj}}) - 1\big)(s_{\mathrm{maj}} - \tfrac{1}{2}).$$

This construction specifies the finite-context averaging forecast in Prediction 3. Its use models the pool through its aggregate rule. The full posterior at finite K is given by the expectation in the proof, which also depends on the joint member agreements and β_K.

Remark 2 (the regime criterion measured on every pool). Using one pool member to summarize selection across targets requires that the same member maximize agreement on every target. Reliable empirical selection also requires separation from each rival at $\sqrt{K}\,\delta/\sigma \gg 1$ (Proposition 2(iii)). Both conditions are measurable before training, and both fail for every pool in this paper at K=16. Across the five real datasets and the five graded structural pools, no cell has one member topping all four evaluation targets ($0/25$). The best member's smallest pairwise separation has median $z = 0.11$. Of the 100 (level, target) cells, 95 lie below 1, and the five exceptions, all in the level-3 degree-selector corner, have $z \leq 1.54$. The complementary mixture arm rotates winners across targets as well, and the best member of the redundant arm has separation $z \leq 0.42$. These separations place the empirical comparisons in the high-error range at $K = 16$. The effective weighting determines how the learner combines the competing members. The mixture experiment in Section 4.1 compares the selection and averaging predictions. The median support size at which the smallest pairwise separation would reach z=2 is 4,610 pairs.

Remark 3 (Within-graph correlation). Let graph g contribute k_g support nodes, with $\sum_{g=1}^m k_g = K$, and write M_g for its number of agreements with the globally aligned hypothesis. Independent demonstration graphs give independent counts M_g, even when votes within a graph are dependent. If $Q_g(\ell) = P(M_g = \ell)$, the probability generating function of the total count is

$$F(z) = \prod_{g=1}^{m} \left(\sum_{\ell=0}^{k_g} Q_g(\ell) z^\ell \right).$$

With $[z^\ell]F(z)$ denoting a polynomial coefficient, the exact polarity recovery probability is

$$P_c^{\text{graph}} = \sum_{\ell > K/2} [z^\ell]F(z) + \frac{1}{2}\mathbf{1}\{K \text{ is even}\}[z^{K/2}]F(z).$$

An independent query with agreement s therefore gives $\Delta_{\text{ICL}} = (2P_c^{\text{graph}} - 1)(s - 1/2)$.

For a mixture of conditionally independent Bernoulli votes with graph agreement S_g, mean $\mathbb{E}S_g = s$, and variance $\tau^2 = \text{Var}(S_g)$, the graph factor is $\mathbb{E}(1 - S_g + S_g z)^{k_g}$. The variance of the total count is

$$\text{Var}\left(\sum_g M_g\right) = Ks(1-s) + \tau^2 \sum_g k_g(k_g - 1).$$

This follows by expanding each count variance into individual variances and pairwise covariances, using $\text{Cov}(A_{gi}, A_{gj}) = \text{Var}(S_g)$ for $i \neq j$. For a polynomial f, expansion around s gives $\mathbb{E}f(S_g) = f(s) + \frac{1}{2}\tau^2 f''(s) + O(\mathbb{E}|S_g - s|^3)$, because the linear term has zero expectation. Thus the leading correction is of second order in the graphwise spread. Class balancing and sampling without replacement can be represented directly through the count law Q_g.

At $K = m$ with one support per independently sampled graph, the votes are independent. At the operating point, the compound-vote calculation calibrated to the per-graph agreement spread of App. D.2 moves the Model A structural efficiency from 0.926 to 0.933 and the R^2 of the zero-parameter surface from 0.891 to 0.893. The real-label experiment with $K = 16$ spread over 2, 4, or 8 demonstration graphs finds no dependence on the number of graphs (paired contrast between 8 and 2 graphs, $t = -0.6$).

B.5 The alignment correspondence

Proposition 3 (The alignment correspondence)**.** *Let $\phi \sim \mathcal{N}(0, \Sigma)$ and consider rules $r(\phi) = \text{sign}(w_r^\top \phi)$ and $e(\phi) = \text{sign}(w_e^\top \phi)$ with positive score variances. Their score correlation is*

$$\rho = \frac{w_r^\top \Sigma w_e}{\sqrt{(w_r^\top \Sigma w_r)(w_e^\top \Sigma w_e)}}.$$

The raw agreement is $1 - \arccos(\rho)/\pi$, and the global polarity-symmetrized agreement is

$$s = \frac{1}{2} + \frac{\arcsin|\rho|}{\pi}.$$

For $\Sigma = cI$ with $c > 0$, ρ is the cosine between w_r and w_e.

Proof. Normalize both scores to unit variance. For $|\rho| < 1$ their joint law can be represented as

$$U = Z_1, \qquad V = \rho Z_1 + \sqrt{1 - \rho^2}\, Z_2,$$

where Z_1, Z_2 are independent standard Gaussian variables. Put $\theta = \arccos\rho$. The two score directions in this plane are $(1, 0)$ and $(\cos\theta, \sin\theta)$, separated by angle θ. The angle of (Z_1, Z_2) is uniform on $[0, 2\pi)$. The two half-plane classifiers disagree on two sectors with total angle 2θ, so

$$P[\text{sign}\, U \neq \text{sign}\, V] = \frac{2\theta}{2\pi} = \frac{\arccos\rho}{\pi}.$$

This is the Gaussian sign-agreement identity (Sheppard, 1899). At $\rho = 1$ the scores agree almost surely, and at $\rho = -1$ they are opposites, giving the same formula at both endpoints. Using $\arcsin\rho + \arccos\rho = \pi/2$ and maximizing the raw agreement over the two global polarities gives $1/2 + |\arcsin\rho|/\pi = 1/2 + \arcsin|\rho|/\pi$. Finally, substituting $\Sigma = cI$ into the covariance formula gives the ordinary weight cosine. □

The zero threshold is the population median of a centered Gaussian score. The testbed uses ranks and empirical graph medians, and its agreement is measured directly under that construction. Under isotropic Gaussian features, agreement is a fixed monotone transform of absolute weight alignment, providing the correspondence with the single-rule alignment analysis of Letey et al. (2026). On the

40-rule structural bank, the isotropic map $\frac{1}{2} + \arcsin|\cos(w_r, w_e)|/\pi$ has Spearman correlation 0.87 with measured s and a mean residual of $+0.10$ (range $+0.02$ to $+0.18$). The covariance formula shows how feature dependence changes the score correlation. Its effect depends on both weight directions and the covariance matrix. Direct label agreement also incorporates the rank transformation and graphwise thresholds used in this bank.

C Robustness of the empirical law

Section 2.2 reports the law on the twelve single-rule cells and summarizes its controls, its statistical support, and its comparison with alternative predictors. This appendix gives the numbers behind those summaries. App. C.1 reports the control battery on every cell, the replication on continuous metrics, the $T{=}8$ pool sweep, and the polarity switch. App. C.2 gives confidence intervals, permutation tests, and the seed replication. App. C.3 compares seven alternative predictors with s, and App. C.4 reports the width sweep.

C.1 Control experiments

Every single-rule run records five controls at evaluation, and a sixth control re-evaluates the checkpoints. App. G defines them and Table 1 reports them for the twelve cells. Zero-shot accuracy on Model A scatters between 0.46 and 0.54 across rules with a mean at chance. On B/C1/C2 it is the constant tie value of Remark 1. Support permutation returns accuracy at $K{=}16$ to 0.49 to 0.52 on average, and the wrong-rule context returns it to 0.49 to 0.51, so the gain requires the correct support labels. Label propagation and the query-feature probe do not involve the model. Both are computed on the fixed evaluation episodes of each stream and are therefore constant across the 40 rules of a ladder. Label propagation scores 0.532 on the structural stream and 0.453 on the feature and mixed streams, and the query-feature probe scores 0.513, 0.510, and 0.502. Neither differs from chance beyond the sampling noise of the fixed evaluation sample. A predictor that receives no information about the query label scores a fixed-sample value that can fall on either side of 0.5.

Convergence screen. A run is classified as unconverged when its final training loss remains within 0.01 of $\ln 2$, and every other run ends at or below 0.66. Four Model B runs stayed at the plateau, two on the structural ladder and two on the feature ladder. Each was retrained with the next seed and the converged run replaces it, so every cell has 40 rules. The support-order control below evaluates the same runs.

Continuous metrics. Replacing Δ_{ICL} by the gain in correct-class probability gives $r = +0.97$ to $+0.99$ over the twelve cells, the gain in log-margin gives $+0.89$ to $+0.98$, and the gain in negative log-likelihood gives -0.81 to -0.95 (Table 1).

The $T{=}8$ pool sweep. The single-rule design is the primary one because a pool of T rules confounds similarity with intra-pool agreement, but the law also holds on pools. Five $T{=}8$ pools, one at each similarity level of the graded bank, trained on Model A with five seeds and evaluated on 1,200 episodes, give $\Delta_{\mathrm{ICL}} = 0.065$, 0.206, 0.335, 0.317, and 0.365 at $s = 0.658$, 0.722, 0.781, 0.841, and 0.894, with $r = +0.921$ over the five levels. In this design the training loss is also lower for the more similar pools, because rules near the evaluation direction are mutually similar. The single-rule cells, in which no intra-pool structure exists, are the controlled version.

The polarity switch. Requirement (3) of Section 2 is tested by a matched pair of T sweeps on Model A ($T \in \{1, 2, 4, 8\}$, five seeds, 1,200 evaluation episodes) that differ only in whether polarity randomization is on. The context gain is defined as accuracy with the correct support minus accuracy with a wrong-rule support at $K{=}16$. With randomization on, zero-shot accuracy is 0.501 and the context gain is $+0.128$, both averaged over T. With randomization off, zero-shot accuracy is 0.633 and rises with T (slope $+0.0060 \pm 0.0034$ on $\log_2 T$, 95% interval), and the context gain is -0.002. A model trained without randomization holds a label convention in its weights and does not use the demonstrations. A model trained with randomization has no such convention and uses them. App. D.4 crosses this switch with pool similarity.

Table 1: Control experiments on the twelve single-rule cells (40 rules each; mean over rules, with the range across rules in brackets). Zero-shot accuracy on B/C1/C2 is the tie value of Remark 1. Support permutation and the wrong-rule context are evaluated at $K{=}16$. The last three columns give the Pearson correlation between s and the $K{=}16$ gain measured in correct-class probability, log-margin, and negative log-likelihood (NLL), with the sign reversed on NLL because it decreases with accuracy.

model	zero-shot	permutation	wrong rule	r prob.	r margin	r NLL
structural ladder						
A	0.491 [0.463,0.523]	0.508 [0.477,0.543]	0.501 [0.462,0.528]	+0.98	+0.90	−0.93
B	0.535	0.513 [0.482,0.550]	0.501 [0.470,0.537]	+0.97	+0.93	−0.87
C1	0.535	0.517 [0.480,0.550]	0.505 [0.473,0.537]	+0.98	+0.92	−0.81
C2	0.535	0.512 [0.482,0.548]	0.504 [0.457,0.543]	+0.98	+0.89	−0.81
feature ladder						
A	0.497 [0.467,0.537]	0.494 [0.455,0.557]	0.494 [0.448,0.550]	+0.99	+0.95	−0.90
B	0.533	0.496 [0.468,0.543]	0.491 [0.447,0.530]	+0.98	+0.98	−0.95
C1	0.533	0.494 [0.462,0.533]	0.496 [0.460,0.535]	+0.99	+0.98	−0.93
C2	0.533	0.499 [0.467,0.528]	0.497 [0.458,0.542]	+0.99	+0.98	−0.87
mixed ladder						
A	0.489 [0.457,0.528]	0.498 [0.465,0.532]	0.506 [0.473,0.537]	+0.98	+0.95	−0.83
B	0.533	0.491 [0.462,0.525]	0.505 [0.457,0.547]	+0.99	+0.98	−0.92
C1	0.533	0.498 [0.470,0.562]	0.504 [0.463,0.533]	+0.99	+0.98	−0.89
C2	0.533	0.498 [0.467,0.532]	0.503 [0.463,0.535]	+0.99	+0.97	−0.85

Table 2: Pearson $r(s, \Delta_{\mathrm{ICL}})$ over the single-rule runs of each cell (40 rules, convergence screen in App. C.1), as printed in Figure 2. Parentheses give the partial correlation controlling for final training loss. Every cell has a bootstrap 95% CI and a permutation $p = 10^{-4}$.

rule family	Model A (GNN proto)	Model B (metagraph)	Model C1 (set tr.)	Model C2 (full tr.)
structural	+0.952 (+0.936)	+0.965 (+0.951)	+0.947 (+0.959)	+0.939 (+0.961)
feature	+0.967 (+0.968)	+0.983 (+0.983)	+0.968 (+0.965)	+0.973 (+0.972)
mixed	+0.936 (+0.900)	+0.964 (+0.936)	+0.957 (+0.934)	+0.955 (+0.931)

Support order. Within a demonstration graph the support nodes are drawn by a class round-robin from a random starting class (Section 2.1). A sixth control fixes the starting class at 0, so the first support node of every graph has label 0. At $K{=}1$ the single support label and at $K{=}4$ all four labels are then 0, whereas at $K{=}16$ the support sets have the same distribution under both rules. Every checkpoint of the twelve cells was evaluated on paired episodes (the same graphs, polarity, and query node; 600 episodes per rule and support size) under the two rules. The paired difference in accuracy is at most 0.010 in any cell at $K \in \{1, 4, 16\}$, so the predictions are stable under this change in support selection.

C.2 Statistical hardening

Every synthetic-ladder correlation in Table 2 is reported with a rule-bootstrap 95% CI (10,000 resamples) and a one-sided permutation p (10,000 permutations). All twelve permutation p values are at the 10^{-4} floor. The tightest interval is feature/Model B at $[+0.973, +0.991]$ and the widest is mixed/Model A at $[+0.894, +0.965]$.

The five-point real-graph correlations receive exact permutation tests over all $5! = 120$ orderings, for which the smallest attainable p is 0.0083. Under the pool-mean axis the structural law gives $p = 0.017$ on Cora and CiteSeer, 0.008 on PubMed, 0.100 on Roman-empire, and 0.383 on Amazon-ratings, so the Amazon-ratings cell is null under that axis. The feature cells all give 0.008. Under the s_{maj} axis the structural cells give $0.008/0.008/0.025/0.117/0.067$. Pooled within-dataset permutation tests on the mean Fisher z reject the joint null at $p = 10^{-4}$ for the structural, the s_{maj}, and the feature laws. The real-label law on Roman-empire has $r = +0.842$ with bootstrap CI $[+0.741, +0.915]$ and $p = 10^{-4}$ ($n = 40$).

Training-seed replication of the Model A structural ladder (identical protocol, seeds 0/1/2, 40 rules each) gives per-seed $r = +0.952/ + 0.955/ + 0.968$. Averaging Δ_{ICL} over seeds before correlating gives $r = +0.978$. The per-rule seed standard deviation of Δ_{ICL} has mean/median/max $= 0.021/0.016/0.065$.

C.3 Predictor baselines

Table 3 compares s with seven alternative predictors on the structural law, and Table 4 repeats the comparison on the real-label tasks. The white-box weight alignment is the cosine between the ground-truth weight vectors of the training rule and the evaluation task, which is the analogue of the alignment scalar of Letey et al. (2026) (Proposition 3). It reaches $r = +0.78$ to $+0.84$ but requires the rules' weights, which only a synthetic testbed supplies. The centroid direction of a task on a graph is the difference between the mean descriptor vectors of its two classes, taken in the space of the three structural statistics on the structural ladder and in the eight-dimensional feature space on real labels. The predictor is the absolute cosine between the centroid directions of the two tasks, averaged over graphs and, on the structural ladder, over the evaluation targets. The linear-probe agreement s_{lin} replaces each task by its linear-probe labeling, the projection of the descriptors onto the task's centroid direction thresholded at the graph median, and is the polarity-symmetrized balanced agreement between the two probe labelings, averaged over graphs. On the real-label tasks the centroid direction and s_{lin} score above s, at $r = +0.899$ and $+0.860$ against $+0.842$.

Table 3: Alternative transfer predictors on the structural law, over the converged single-rule runs, as Pearson r / LOO-Q^2. LOO-Q^2 uses a linear fit with a free slope and intercept for every predictor, not the fixed-intercept η form of Table 9. A-priori predictors need no trained model, and post-hoc predictors use the trained checkpoint and target data. The label-marginal match is exactly constant because median-threshold rules are balanced by construction. LEEP is exactly constant on B/C1/C2 because their tied $K{=}0$ logits give every source model the same uniform predictive (Corollary 1).

predictor	type	Model A	Model B	Model C1	Model C2
s (ours)	a-priori	$\mathbf{.952/.90}$	$\mathbf{.965/.92}$	$\mathbf{.947/.89}$	$\mathbf{.939/.87}$
raw agreement	a-priori	$.177/-.06$	$.135/-.08$	$.180/-.07$	$.140/-.08$
weight alignment	a-priori	$.777/.56$	$.839/.67$	$.810/.61$	$.813/.62$
centroid direction	a-priori	$.747/.47$	$.758/.49$	$.763/.49$	$.753/.48$
label marginal	a-priori	exactly constant			
LogME	post-hoc	$.924/.84$	$.911/.81$	$.734/.50$	$.798/.61$
LEEP ($K{=}0$)	post-hoc	$.503/.18$	exactly constant		
train loss	post-hoc	$.551/.25$	$.531/.22$	$.407/.11$	$.378/.07$

Table 4: The same comparison on the real-label tasks (Roman-empire, $n = 40$), as r and LOO-Q^2 under a free linear fit rather than the η form of Table 9. The white-box weight alignment is undefined for class-subset tasks.

predictor	r	LOO-Q^2
balanced agreement s (real-label axis)	$+0.842$	$+0.684$
linear-probe agreement s_{lin}	$+0.860$	$+0.708$
centroid direction	$\mathbf{+0.899}$	$\mathbf{+0.790}$
raw balanced agreement	$+0.204$	-0.090
label marginal	$+0.146$	-0.108
LogME (post-hoc)	$+0.765$	$+0.556$
LEEP (post-hoc, $K{=}0$)	$+0.460$	$+0.155$
weight alignment	undefined	

C.4 Width sweep

The structural law was re-run at hidden widths 32, 64, 128, and 256 on Model A and at the same model widths on Model C1, with 40 rules per width, the single-rule protocol of Section 2.1, seed 0, and the fixed 15-epoch budget. Width 64 is the main-text configuration. The parameter count

runs from 10,369 to 598,817 on Model A and from 17,665 to 1,058,817 on Model C1, a $58\times$ to $60\times$ range. Table 5 reports the correlation, the finite-K-corrected efficiency, and the R^2 of the zero-parameter surface at each width. On Model A the correlation varies by at most 0.015 across widths, and the efficiency lies between 0.94 and 1.00 with its largest value at the largest width. On Model C1 the correlation stays at or above $+0.831$ at every width, the largest width loses efficiency, and the two largest widths lose surface fit under the fixed budget. A compute-matched control retrains the width-256 Model C1 for 30 epochs. Its efficiency increases from 0.741 to 0.767 and remains below the smaller models at this data budget.

Table 5: The structural law across model width (40 rules per cell, the single-rule protocol of Section 2.1). $\eta_{\mathrm{fin}K}$ is the finite-K-corrected efficiency and F1 R^2 the fit of the zero-parameter surface. The bottom row is the compute-matched control at double the training budget.

	Model A			Model C1		
width	r	$\eta_{\mathrm{fin}K}$	F1 R^2	r	$\eta_{\mathrm{fin}K}$	F1 R^2
32	$+0.962$	0.967	0.888	$+0.945$	0.812	0.823
64	$+0.952$	0.941	0.891	$+0.947$	0.790	0.807
128	$+0.966$	0.979	0.899	$+0.879$	0.802	0.659
256	$+0.952$	0.998	0.917	$+0.831$	0.741	0.505
256, $2\times$ budget				$+0.863$	0.767	

D TESTING THE THREE PREDICTIONS

Section 4.1 tests Predictions 1 to 3 and shows one panel per prediction in Figure 3. This appendix reports the complete fit ladder of the finite-K surface (App. D.1), derives the intercept of the unconstrained line (App. D.2), reports the two diversity axes in full (App. D.3) and the polarity grid (App. D.4), and reports the mixture experiment with its trainability controls (App. D.5).

D.1 THE FINITE-K SURFACE IN ALL CELLS

Table 6: Fit ladder for the finite-K law over all (rule, K) points of each cell. F0 applies the asymptote $s - \frac{1}{2}$ at every K. F1 uses the zero-parameter surface of Theorem 2, and F2 multiplies that surface by one fitted efficiency η. The one-shot exponent is fitted on $(s - \frac{1}{2})$, compared with 2 for the single-target law and 1 for the asymptote. The final column gives asymptotic and finite-K-corrected efficiencies at $K = 16$. B/C1/C2 use $\mathrm{acc}(K) - \frac{1}{2}$ to account for the $K = 0$ tie offset.

cell	F0 R^2	F1 R^2	F2 R^2 (η)	K=1 pow	collapse slope	$\eta \uparrow$
struct/A	0.31	0.89	0.93 (.93)	1.92	0.94	.91 → .93
struct/B	0.28	0.88	0.92 (.91)	2.21	0.92	.87 → .89
struct/C1	0.21	0.81	0.86 (.92)	1.64	0.95	.88 → .90
struct/C2	0.25	0.84	0.87 (.93)	1.68	0.97	.91 → .93
feat/A	0.56	0.92	0.93 (.96)	1.53	0.96	.91 → .95
feat/B	0.58	0.92	0.93 (.96)	1.47	0.98	.89 → .94
feat/C1	0.56	0.91	0.92 (.95)	1.24	1.00	.88 → .92
feat/C2	0.57	0.92	0.93 (.95)	1.51	0.99	.89 → .93
mixed/A	-0.16	0.79	0.89 (.84)	1.24	0.85	.78 → .84
mixed/B	-0.42	0.76	0.93 (.79)	1.50	0.78	.74 → .80
mixed/C1	-0.45	0.77	0.94 (.80)	1.20	0.82	.76 → .81
mixed/C2	-0.55	0.73	0.92 (.79)	1.39	0.82	.75 → .81

Table 6 fits the stored curves at $K \in \{1, 4, 8, 16\}$ in every cell with three forms, namely the asymptote applied at every K (F0), the zero-parameter surface of Theorem 2 (F1), and the surface times one fitted efficiency (F2). Model B, Model C1, and Model C2 are scored against the theory baseline $\mathrm{acc}(K) - \frac{1}{2}$ because their K=0 intercept includes the tie offset of Remark 1. On the feature ladder the parameter-free asymptote reaches $R^2 = 0.926$ on Model A (structural ladder 0.801, Section 4.1).

On the mixed ladder and on Model B, Model C1, and Model C2, the fitted η form holds at $\eta = 0.79$ to 0.96 (column F2). F0 fits $R^2 = 0.21$ to 0.58 on the structural and feature ladders and gives negative values on the mixed ladders, so the finite-context vote factor is needed. Accounting for that factor increases every fitted efficiency (column $\eta \uparrow$; Model A structural $0.91 \to 0.93$). The mixed cells have the lowest F1 values. Lemma 1 gives the exact dependence of finite-context efficiency on rule error and output error, including the change in polarity recovery caused by rule error.

At $K{=}1$ the surface reduces to $\Delta_{\mathrm{ICL}}(1, s) = 2(s - \frac{1}{2})^2$ (Figure 5). The fitted exponents on the four structural ladders are 1.92, 2.21, 1.64, and 1.68 against the predicted 2, and the exponent 1 of the asymptote fits at $R^2 < -2$. The feature-ladder exponents are lower, 1.2 to 1.5. Averaging across targets changes the fitted exponent on the structural ladder. The registered similarity averages the agreement over the evaluation targets (rule 10 and the three latents of rule 13), and the one-shot gain is convex in the agreement. On that averaged axis the reference learner's own one-shot gains follow an exponent of 1.82 on the structural ladder and 1.99 on the feature ladder. The bootstrap intervals of Model A, Model C1 and Model C2 contain 1.82, and those of Model A and Model B also contain 2. The exponents below these values on the feature and mixed ladders are a property of the learners at $K{=}1$.

The surface of Theorem 2 is stated per evaluation target. The ladder's axis averages the agreement over the targets, and the reference learner's prediction on this axis is the average of the per-target surfaces. Applying the surface to the averaged agreement is an approximation. For target weights $\alpha_t \geq 0$ with $\sum_t \alpha_t = 1$, put $\bar{s} = \sum_t \alpha_t s_t$. At one shot the difference is exact,

$$\sum_t \alpha_t f_1(s_t) - f_1(\bar{s}) = 2\left(\sum_t \alpha_t s_t^2 - \bar{s}^2\right) = 2\,\mathrm{Var}_\alpha(s_t).$$

Here $f_K(s)$ denotes the surface of Theorem 2. As $K \to \infty$, $f_K(s) \to s - 1/2$, so averaging commutes with the limit. On the structural ladder the two differ by 0.016 on average at $K{=}1$ (at most 0.027) and by at most 0.009 at $K \geq 8$, and on the feature and mixed ladders by less than 0.010 at every K. The tables use the approximation. On the Model A structural cell its F1 R^2 is 0.891 against 0.869 for the averaged surface.

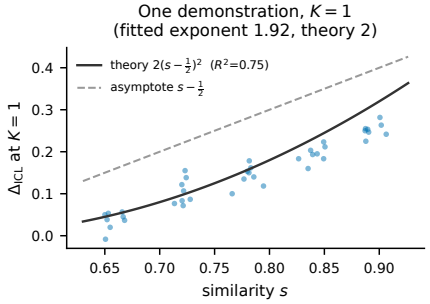

Figure 5: The one-shot law. Δ_{ICL} at $K{=}1$ against s on the Model A structural ladder, with the predicted $2(s - \frac{1}{2})^2$ and the asymptote $s - \frac{1}{2}$. The fitted exponent is 1.92, between the reference learner's 1.82 on this axis and 2.

D.2 The intercept of the unconstrained line

For a fixed task pair, let m_G be its raw agreement on graph G and put $d_G = m_G - 1/2$. The measured graphwise similarity and global agreement margin are

$$s = \frac{1}{2} + \mathbb{E}_G|d_G|, \qquad \mu = |\mathbb{E}_G d_G|.$$

Choose a global sign $\omega \in \{-1, +1\}$ such that $\mathbb{E}_G[\omega d_G] = \mu$. The folding excess has the exact form

$$(s - \tfrac{1}{2}) - \mu = \mathbb{E}_G\big(|\omega d_G| - \omega d_G\big) = 2\mathbb{E}_G(-\omega d_G)_+ \geq 0,$$

where $(x)_+ = \max(x, 0)$. The excess is twice the mean magnitude of negative agreement margins under the chosen global orientation. The independent-vote law uses the global similarity $1/2 + \mu$,

giving

$$\Delta_{\mathrm{ICL}}(K) = \big(2P_c(K, \tfrac{1}{2} + \mu) - 1\big)\mu.$$

The correction therefore enters both the vote factor and the query agreement margin.

Let x_i denote a ladder's measured similarities and y_i the corresponding theoretical gains. The unconstrained least-squares line $y = ax + b$ has

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \qquad b = \bar{y} - a\bar{x}, \qquad x_0 = -\frac{b}{a} = \bar{x} - \frac{\bar{y}}{a},$$

for nonzero a. Multiplying every theoretical gain by a constant efficiency multiplies a and b by the same constant and leaves x_0 unchanged.

The unconstrained line on the Model A structural ladder has $x_0 = 0.571$. Fitting a line to the finite-K surface at $K = 16$ on the ladder's measured x_i gives $x_0 = 0.542$. This displacement arises from the curvature of the vote factor at low similarity. Re-measuring raw per-graph agreements gives folding excesses $0.026/0.013/0.004/0.002/0.000$ from the lowest to highest level. Using $1/2 + \mu$ in the theoretical gains moves the predicted intercept to 0.565. Both calculations evaluate the theoretical curve at the measured similarities, with no fitted efficiency.

D.3 Diversity nulls

Family diversity (M). Per-seed OLS slopes of Δ_{ICL} over $M \in \{1, 2, 4, 8\}$ graph families (four training rules and the held-out tasks fixed; 5 seeds; 1,200 evaluation episodes; mean $\pm$ 2SE) are $+0.0010 \pm 0.0034$ on Model A, -0.0008 ± 0.0048 on Model B, -0.0060 ± 0.0034 on Model C1, and -0.0057 ± 0.0051 on Model C2. No model rises with M. The two transformers show a small negative slope whose interval excludes zero, which is the opposite sign of a diversity-driven increase. Per-M curves are in Figure 6. The training prior adds families in a fixed order, ER, SBM, BA, WS, then a motif generator that grows a short path by attaching motifs drawn from a bank of triangles, four-cliques, stars, houses, paths, and cycles, then random labeled trees, two-dimensional grids with $\lfloor\sqrt{n}\rfloor$ rows, and rings. At M=1 the prior uses ER alone and at M=8 it uses all eight families.

Rule count (T) under coverage control. Every member of the two pools of Section 4.1 uses all three structural statistics of the evaluation rules, degree, clustering coefficient, and triangle count, so predicate coverage is saturated at T=1 and constant in T. In the similar pool each member is the balanced majority of the three statistics and one or two feature coordinates (rule ids 48 to 63). In the dissimilar pool each member is a linear threshold rule over the three statistics with signed integer weights of mixed sign, the 16 covering weight vectors with the lowest agreement with the evaluation rules under a grid search (rule ids 64 to 79). The two pools differ in their measured similarity to the evaluation rules (5 seeds; 1,200 evaluation episodes). At T=16 the similar pool gives $\Delta_{\mathrm{ICL}} = 0.346 \pm 0.007$ on Model A and 0.343 ± 0.012 on Model B, and the dissimilar pool gives 0.107 ± 0.028 and 0.064 ± 0.056. Both arms are flat in T (Figure 3c). The gain tracks the pool's agreement with the evaluation rules and remains stable across the tested rule counts.

In the cell most favorable to a diversity-driven transition from in-weights to in-context behavior, polarity randomization off with a dissimilar pool, the gain does not increase with T in the polarity $\times$ similarity grid of App. D.4 (slope $+0.012$ on $\log_2 T$, interval including zero). Within the tested range, rule-count diversity does not open an in-context channel in this regime, consistent with alignment-based accounts (Letey et al., 2026).

D.4 The polarity $\times$ similarity grid

App. C.1 reports the polarity switch on a T sweep whose evaluation rules are similar to the training pool. Table 7 crosses the switch with pool similarity. The similar pool is the graded level 0.894 (rule ids 112 to 119) and the dissimilar pool is the level 0.658 (ids 80 to 87). Both are evaluated on the fixed held-out tasks at $T \in \{1, 2, 4, 8\}$ with three seeds and 500 evaluation episodes. The table reports zero-shot accuracy and the context gain (accuracy with the correct support minus accuracy with a wrong-rule support at K=16), both averaged over T, and the per-seed OLS slope of the context gain on $\log_2 T$ (mean $\pm$ 2SE).

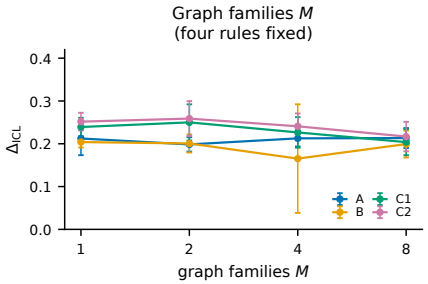

Figure 6: Graph-family diversity M (mean over 5 seeds with 95% CI). No model gains with M.

Table 7: The polarity $\times$ similarity grid (three seeds, values averaged over $T \in \{1, 2, 4, 8\}$).

cell	acc_0	context gain	gain slope on $\log_2 T$
similar $\times$ ON	0.494	$+\mathbf{0.370}$	-0.0064 ± 0.0070
similar $\times$ OFF	0.740	-0.004	-0.0041 ± 0.0096
dissimilar $\times$ ON	0.499	$+0.107$	-0.0103 ± 0.0073
dissimilar $\times$ OFF	0.548	$+0.007$	$+0.0117 \pm 0.0165$ (n.s.)

D.5 The mixture experiment

The three pools were selected from the same integer-weight family as the graded bank, matched on individual similarity, and differing in error geometry. The predictions were computed from the measured pool statistics with the single-rule efficiency of 0.94 recorded at pre-registration and the finite-K factor at K=16, and they were recorded before training (pre-registration in the released repository). All arms use Model A and the protocol of the T=1 law runs, with five seeds for the C and R arms, three seeds for the S arm and the two budget controls, and 600 evaluation episodes.

Table 8: The pre-registered mixture experiment. Predictions computed from measured pool statistics with $\eta = 0.94$ and the finite-K factor at K=16, with both Bayes endpoints shown. The discriminating contrast is $\Delta^C_{\text{ICL}} - \Delta^R_{\text{ICL}}$, predicted as -0.006 under selection and $+0.047$ under averaging, with measured value $+\mathbf{0.049}$ from unrounded seed means (5v5 seeds fully separated, Welch $t \approx 4.2$, $p \approx 0.003$).

arm	T	s_j	s_{maj}	pred. (selection)	pred. (averaging)	measured Δ_{ICL}
C complementary	3	≈ 0.79	0.840	0.274	$\mathbf{0.319}$	$0.297 \pm .019$
R redundant	3	≈ 0.79	0.793	0.280	0.272	$0.249 \pm .017$
S single	1	0.801	—	0.280	0.280	$0.276 \pm .011$
C, half training budget	3	control, 8 epochs				$0.301 \pm .009$
R, double training budget	3	control, 30 epochs				$0.266 \pm .012$

The gain of arm C remains stable when its training budget is halved. Doubling the budget of arm R brings its gain close to its averaging prediction, below the value of C. Arm C lies between the two endpoints ($0.297 \in [0.274, 0.319]$), supporting the averaging prediction for the complementary pool. Both pooled arms lie 0.02 to 0.03 below their absolute predictions, which reuse the pre-registered single-rule efficiency of 0.94, while the single-rule arm matches its prediction to 0.004. The pre-registered contrast is differenced against this common shift. Diversity therefore raises the gain when it raises s_{maj}, and the pre-registered amount matches. The M and T sweeps of App. D.3 could not raise s_{maj} by construction, so the diversity nulls and this positive effect follow from the same posterior.

E TRANSFER ON REAL GRAPHS

Section 4.2 reports out-of-sample forecasts, the real-structure and real-feature laws on five datasets, the real-label law, and the identifiability screen. This appendix states the real-graph protocol (App. E.1), the forecast protocol and scores (App. E.2), the per-dataset tables (App. E.3), the errors-in-variables accounting of every real-graph cell (App. E.4), and the identifiability screen with its measurements (App. E.5).

E.1 REAL-GRAPH PROTOCOL

Episode graphs. Each episode graph is a connected subgraph of one real network, grown by breadth-first search from a random seed node with a randomized frontier order, with the node count drawn as for the synthetic families (32 to 64 nodes). Each subgraph is a separate graph object with its own node indexing and no edges to any other subgraph of the episode, so the disconnection requirement holds regardless of node overlap between subgraphs. Subgraphs of one episode may share source nodes. For structural and feature rules the label of a node is defined relative to its subgraph, so a shared node carries no label information across subgraphs. For real-label tasks the query node lies among the $K{=}16$ support nodes with probability at most K/N, below 0.6% on Cora and CiteSeer and below 0.1% on Roman-empire and Amazon-ratings. The datasets are Cora, CiteSeer, and PubMed (Yang et al., 2016) and the heterophilous Roman-empire and Amazon-ratings (Platonov et al., 2023). Episodes keep $m{=}4$ demonstration graphs, one query graph, K drawn from $\{0, 1, 4, 8, 16\}$ in training, and polarity randomization.

Three settings. The real-structure setting keeps the synthetic i.i.d. features, so the structural rules are unchanged and only the degree, clustering, and triangle distributions are real. The real-feature setting replaces the features by the leading eight principal components of the dataset's node features, standardized per graph, so the feature ladder runs unchanged with $d{=}8$. The real-label setting replaces the rule by a class-subset dichotomy of the dataset's node classes, on real features.

Training pools and runs. The real-structure and real-feature runs train Model A on the five graded pools of eight rules each (structural rule ids 80 to 119; feature rule ids 124 to 163) and evaluate on the fixed held-out tasks of the corresponding ladder, with three seeds per level. The real-structure runs use 500 evaluation episodes per support size and the real-feature runs 1,200 (Table 14). All isolation controls run on every real-graph cell. Label propagation and zero-shot accuracy are at chance, and support permutation and the wrong-rule context remove the gain.

Re-measured axis. The registered similarities were measured on synthetic graphs. Real degree, clustering, and triangle distributions change the per-rule agreements, so the synthetic values would misplace every point on the x axis. Each level's agreement is therefore re-measured on subgraphs of the target dataset with the protocol that defined the synthetic axis, namely per-graph agreement, polarity-symmetrized as $\max(a, 1-a)$, with selector rules averaged over their latents. The structural axis uses 200 subgraphs of 48 nodes per dataset and the feature axis 300 subgraphs. Under polarity randomization an anti-aligned rule transfers exactly like an aligned one, so the symmetrized axis is the quantity the theory refers to. The structural cells are reported under two pool statistics, the registered default (the mean of the member agreements) and the posterior-vote statistic s_{maj} of Proposition 2. App. E.4 gives both for all five datasets.

Real-label tasks. A task labels node v by $\mathbf{1}[\mathrm{class}(v) \in S]$ for a subset S of the dataset's classes. Under polarity randomization a subset and its complement are the same task, so task similarity is the label agreement with the reference task, symmetrized as $\max(a, 1-a)$, balanced with respect to the reference task to remove class imbalance, and measured on 300 real-feature subgraphs. On Roman-empire (18 classes) the reference task is the subset $\{0, \ldots, 8\}$, and the bank is 40 size-9 subsets at graded overlap with the reference, whose measured similarity spans 0.57 to 0.95. Each run trains Model A on one task ($T{=}1$) with three seeds and 600 evaluation episodes and is evaluated on the reference task. The other banks enumerate the admissible subsets. CiteSeer uses all size-2 and size-3 subsets of its 6 classes minus the reference dichotomy and its complement ($35 - 2 = 33$ tasks), Cora all size-3 subsets of 7 classes minus the reference ($35 - 1 = 34$), and Amazon-ratings

all nontrivial subsets of 5 classes minus the reference and its complement ($30 - 2 = 28$). PubMed has 3 classes and no class-subset ladder.

Efficiency convention. Two efficiencies are fitted to a real-graph law, both by least squares through the origin. The asymptotic efficiency fits $\Delta_{\mathrm{ICL}} = \eta\,(s - \frac{1}{2})$. The finite-$K$ efficiency fits $\Delta_{\mathrm{ICL}} = \eta\, f_{16}(s)$ with f_{16} the surface of Theorem 2 at K=16, with Δ_i the gain of task i. Differentiating $\sum_i(\Delta_i - \eta f_i)^2$ gives the normal equation $\sum_i f_i\Delta_i = \hat{\eta}\sum_i f_i^2$, hence $\hat{\eta} = \sum_i f_i\Delta_i / \sum_i f_i^2$ when $\sum_i f_i^2 > 0$. Efficiencies on real graphs are finite-K efficiencies unless stated otherwise. On the real-label law of Roman-empire the asymptotic value is 0.43 and the finite-K value is 0.45. A line with a free intercept has slope 0.52 and crosses zero at the measured $s = 0.554$, against the 0.549 that App. E.5 predicts from the symmetrization floor. Importing η from the synthetic or real-feature settings gives $Q^2 < 0$.

E.2 Out-of-sample forecasts

Table 9 scores the law as a forecaster. The score is Q^2, the out-of-sample R^2 computed from the prediction residuals, and the forecast mean absolute error (MAE) is in parentheses. On the synthetic ladders the protocol is leave-one-direction-out. The efficiency is fitted on the other weight directions and the gains of the held-out direction are forecast from similarity alone. Leave-one-rule-out scores differ from those reported by at most 0.003.

On the real-feature ladder the protocol is leave-one-dataset-out. The efficiency is fitted on the 20 level points of four datasets, the five points of the fifth are forecast, and the pooled score over the 25 held-out level points is $Q^2 = 0.877$. The same protocol on the real-structure ladder gives pooled $Q^2 = -0.032$ under the pool-mean axis and 0.339 under the s_{maj} axis. Within each dataset the gain tracks similarity, but each dataset has its own efficiency, from 0.64 to 0.91 under s_{maj}, so an efficiency fitted on the other four misplaces the level, most on PubMed (forecast MAE 0.090). The synthetic-to-real forecast uses $\eta = 0.976$ from the synthetic feature ladder only and forecasts all 25 real points with no real-graph fitting. The real-label forecast is leave-one-task-out over the 40 Roman-empire tasks.

Table 9: Out-of-sample forecast scores, Q^2 with the forecast mean absolute error in parentheses.

leave-one-direction-out	Model A	Model B	Model C1	Model C2
structural	0.849 (.030)	0.791 (.038)	0.810 (.032)	0.821 (.030)
feature	0.925 (.025)	0.906 (.024)	0.907 (.025)	0.906 (.027)
mixed	0.847 (.026)	0.761 (.030)	0.771 (.027)	0.814 (.023)
no fitted parameter (Model A)	**structural**	**feature**		
asymptote $s - \frac{1}{2}$	0.801	0.926		
finite-K surface	**0.885**	0.911		
real graphs	**Q^2 (MAE)**			
real features, leave-one-dataset-out	0.877 (.031)			
real structure, pool-mean axis	−0.032 (.054)			
real structure, s_{maj} axis	0.339 (.043)			
synthetic to real	0.872 (.031)			
real labels, leave-one-task-out	0.661 (.041)			

E.3 Per-dataset results

Table 10 lists every level point behind Figure 4a and the per-dataset correlations of the structural and feature laws. On the homophilous graphs the structural law follows a local inversion of the axis. Cora's levels 0 and 1 nearly tie in measured similarity (0.732 and 0.725) and tie in Δ_{ICL} within error (0.251 and 0.263).

Table 10: Per-level real-graph results (3 seeds, T=8, mean ± 95% CI; App. E.1). The left block gives the structural ladder on real structure with similarity re-measured under the builder protocol. The right block gives the feature ladder on real PCA features. The last row of each block gives the per-dataset correlation.

		structural (real structure)		feature (real features)	
dataset	level	s	Δ_{ICL}	s	Δ_{ICL}
Cora	L0	0.732	0.251 ± 0.058	0.598	0.024 ± 0.042
	L1	0.725	0.263 ± 0.170	0.650	0.114 ± 0.074
	L2	0.869	0.355 ± 0.042	0.726	0.238 ± 0.046
	L3	0.907	0.368 ± 0.026	0.789	0.261 ± 0.076
	L4	0.936	0.383 ± 0.025	0.862	0.386 ± 0.056
$r(s, \Delta_{\mathrm{ICL}})$			$+0.992$		$+0.987$
CiteSeer	L0	0.702	0.219 ± 0.028	0.585	0.031 ± 0.038
	L1	0.693	0.230 ± 0.112	0.644	0.125 ± 0.029
	L2	0.835	0.315 ± 0.032	0.704	0.204 ± 0.035
	L3	0.908	0.346 ± 0.081	0.770	0.251 ± 0.010
	L4	0.925	0.356 ± 0.035	0.850	0.351 ± 0.025
$r(s, \Delta_{\mathrm{ICL}})$			$+0.993$		$+0.993$
PubMed	L0	0.766	0.181 ± 0.026	0.643	0.022 ± 0.040
	L1	0.757	0.177 ± 0.016	0.706	0.208 ± 0.053
	L2	0.866	0.270 ± 0.023	0.756	0.278 ± 0.040
	L3	0.920	0.276 ± 0.044	0.823	0.329 ± 0.013
	L4	0.940	0.299 ± 0.009	0.891	0.365 ± 0.028
$r(s, \Delta_{\mathrm{ICL}})$			$+0.981$		$+0.932$
Roman-empire	L0	0.647	0.119 ± 0.047	0.591	0.025 ± 0.093
	L1	0.709	0.272 ± 0.021	0.613	0.049 ± 0.043
	L2	0.772	0.321 ± 0.019	0.663	0.161 ± 0.027
	L3	0.821	0.292 ± 0.016	0.753	0.220 ± 0.048
	L4	0.886	0.291 ± 0.057	0.814	0.278 ± 0.042
$r(s, \Delta_{\mathrm{ICL}})$			$+0.732$		$+0.977$
Amazon-ratings	L0	0.711	0.211 ± 0.071	0.600	0.025 ± 0.074
	L1	0.734	0.322 ± 0.063	0.661	0.108 ± 0.035
	L2	0.745	0.350 ± 0.029	0.710	0.197 ± 0.074
	L3	0.748	0.205 ± 0.103	0.767	0.258 ± 0.018
	L4	0.862	0.316 ± 0.034	0.836	0.326 ± 0.078
$r(s, \Delta_{\mathrm{ICL}})$			$+0.368$		$+0.993$

E.4 Boundary accounting

Table 11 asks whether each observed five-point correlation is consistent with the law under measurement noise. For every real-graph cell a parametric bootstrap draws the x noise from a graph bootstrap of the similarity measurement and the y noise from the seed spread, with η fitted leave-that-dataset-out, and yields a prediction interval for the observed correlation. The structural cells are evaluated under both the pool-mean axis and the s_{maj} axis. Under the pool-mean axis the homophilous structural cells and all feature cells lie inside or just below their intervals, whereas the two heterophilous structural cells lie below theirs, so these two cells require a different pool statistic to describe the association. The averaging-regime statistic substantially increases their correlations. Under the s_{maj} axis Amazon-ratings moves inside its interval ($+0.904$), whereas the selection statistic $\max_j s_j$ gives $+0.241$. The re-measured axis is applied to all five datasets. Roman-empire rises from $+0.732$ to $+0.860$, PubMed decreases from $+0.981$ to $+0.850$, and Cora and CiteSeer move from $+0.992$ and $+0.993$ to $+0.967$ and $+0.980$. PubMed and Roman-empire retain bounded residuals under both axes that noise does not explain. The three axes were fixed before this analysis, since the pool mean is the registered default, s_{maj} and $\max_j s_j$ are the two pool statistics that Section 3 derives, and the pre-registered mixture experiment (App. D.5) independently identifies averaging as the realized regime.

Table 11: Errors-in-variables accounting of every real-graph cell, with parametric-bootstrap prediction intervals for the observed five-point correlation (graph-bootstrap x noise, seed y noise, η fitted leave-that-dataset-out). The structural cells are shown under both the pool-mean x and the posterior-vote s_{maj} x of Proposition 2. A cell is marked far below when the observed correlation falls more than 0.2 under its interval.

axis	dataset	r observed	predicted interval	verdict
structural, pool mean	Cora	$+0.992$	$[+0.91, +1.00]$	inside
	CiteSeer	$+0.993$	$[+0.97, +1.00]$	inside
	PubMed	$+0.981$	$[+0.99, +1.00]$	below
	Roman-empire	$+0.732$	$[+0.99, +1.00]$	**far below**
	Amazon-ratings	$+0.368$	$[+0.87, +1.00]$	**far below**
structural, s_{maj}	Cora	$+0.967$	$[+0.84, +1.00]$	inside
	CiteSeer	$+0.980$	$[+0.94, +1.00]$	inside
	PubMed	$+0.850$	$[+0.98, +1.00]$	below
	Roman-empire	$+0.860$	$[+0.98, +1.00]$	below
	Amazon-ratings	$+\mathbf{0.904}$	$[+0.77, +1.00]$	**inside**
feature	Cora	$+0.987$	$[+0.97, +1.00]$	inside
	CiteSeer	$+0.993$	$[+0.99, +1.00]$	inside
	PubMed	$+0.932$	$[+0.99, +1.00]$	below
	Roman-empire	$+0.977$	$[+0.97, +1.00]$	inside
	Amazon-ratings	$+0.993$	$[+0.97, +1.00]$	inside

E.5 Identifiability of real-label tasks

The law of Section 3 presupposes a rule that the demonstration graphs and the query graph share. Synthetic rules satisfy this by construction, and a dataset's own class labels need not, because the relation between node features and class membership can differ from subgraph to subgraph. This section measures that relation before training. A linear probe is fitted to the reference task on each real-feature subgraph, and the consistency of the probe directions across subgraphs is summarized in two ways. The screen, the mean absolute cosine between per-subgraph directions, is the scalar from which the outcomes on Amazon-ratings and CiteSeer were pre-registered. The invariant mass $\hat{\lambda}^2$, the squared norm of the mean direction corrected for probe estimation error (Proposition 4), is the derived quantity used in the analysis below. A task with no invariant component leaves no shared rule for the learner to recover, and its minimum training loss is $\ln 2$ (Proposition 4(ii)). Table 12 lists both quantities per dataset with the real-label outcomes, including the two confirmed predictions and the Cora test.

Table 12: Identifiability of real-label tasks. "Screen" is the pre-registration scalar, the mean $|\cos|$ between per-subgraph probe directions, against an 8-D random baseline of ≈ 0.29. $\hat{\lambda}^2$ is the derived invariant mass of Proposition 4, a signed consistency corrected for attenuation, with bootstrap CIs over graphs. CiteSeer and Amazon-ratings outcomes were pre-registered from the screen.

dataset	screen	$\hat{\lambda}^2$ [CI]	real-label law	note
Roman-empire	0.437	0.864 [.80, .96]	$r = +0.842$, $\eta = 0.45$	reference bank
CiteSeer	0.410	0.532 [.47, .63]	$r = +0.628$, $\eta = 0.45$	pre-registered, confirmed
PubMed	0.352	0.334 [.27, .43]	—	3 classes, no subset ladder
Cora	0.310	0.085 [.05, .20]	$r = +0.821$, $\eta = 0.36$	screen prediction falsified
Amazon-ratings	0.298	0.060 [.03, .15]	flat (loss at $\ln 2$)	pre-registered, confirmed

Model. For this analysis, model the node observables as $\phi \sim \mathcal{N}(0, I_d)$, with $d \geq 2$, and labels as $Y = \varsigma \operatorname{sign}\langle u_g, \phi \rangle$. The directions u_g are unit vectors drawn independently across graphs, and ς is the episode polarity. Define the invariant mass by $\lambda^2 = \|\mathbb{E} u_g\|^2$. The unconditional model, denoted (M-u), specifies

$$u_g = \lambda w + \sqrt{1 - \lambda^2}\, \xi_g, \qquad \|w\| = 1, \quad 0 \leq \lambda \leq 1,$$

where ξ_g is uniform on the unit sphere of $w^\perp$. Conditional on the task parameters, these residual directions are independent across graphs and independent of all label-free observables. Thus the query residual is independent of the demonstrations, the query features, and the query graph structure. The Gaussian assumption defines the analytical feature model. The empirical features are the standardized eight-dimensional inputs of App. E.1.

Proposition 4 (Invariant mass and symmetry)**.** (i) *For independent graph directions,* $\mathbb{E}\langle u_g, u_{g'}\rangle = \lambda^2$*. Suppose unit, label-oriented probes based on* n *labels satisfy* $\mathbb{E}[\hat{w}_{g,n} \mid u_g] = b_n u_g$*, with* $b_n > 0$ *common across graphs and with conditionally independent estimation errors. Their signed cross-graph consistency is*

$$\rho_{\mathrm{cross},n} := \mathbb{E}\langle \hat{w}_{g,n}, \hat{w}_{g',n}\rangle = b_n^2 \lambda^2.$$

If the probe reliability additionally satisfies $b_n^{-2} - 1 = v/n$ *for a fixed* $v \geq 0$*, two independent half-sample probes identify the full-sample reliability through*

$$\rho_{\mathrm{half}} = b_{n/2}^2, \qquad b_n^2 = \frac{2\rho_{\mathrm{half}}}{1+\rho_{\mathrm{half}}}, \qquad \lambda^2 = \frac{\rho_{\mathrm{cross},n}}{b_n^2}.$$

(ii) *Let* $\mathcal{F}$ *contain the entire observed episode and write* $U_q = \varsigma u_{g_q}$ *for the label-oriented query direction. If* $\mathcal{L}(U_q \mid \mathcal{F}) = \mathcal{L}(-U_q \mid \mathcal{F})$ *and the query score is nonzero almost surely, every predictor based on* $\mathcal{F}$ *has accuracy* $1/2$*. The minimum expected cross-entropy is* $\ln 2$*. Model (M-u) satisfies this condition at* $\lambda = 0$*.*

Proof. For *(i)*, independence and bilinearity give

$$\mathbb{E}\langle u_g, u_{g'}\rangle = \sum_{\ell=1}^{d} \mathbb{E}[u_{g,\ell}]\mathbb{E}[u_{g',\ell}] = \|\mathbb{E}u_g\|^2.$$

Applying conditional independence to the probes gives

$$\mathbb{E}\langle \hat{w}_{g,n}, \hat{w}_{g',n}\rangle = \mathbb{E}\langle b_n u_g, b_n u_{g'}\rangle = b_n^2 \lambda^2.$$

For two probes fitted on independent halves of the same graph sample, conditioning on the common direction gives $\mathbb{E}[\langle \hat{w}^{(1)}_{g,n/2}, \hat{w}^{(2)}_{g,n/2}\rangle \mid u_g] = b_{n/2}^2$. The reliability law then implies

$$\rho_{\mathrm{half}}^{-1} - 1 = \frac{2v}{n}, \qquad b_n^{-2} - 1 = \frac{v}{n} = \frac{\rho_{\mathrm{half}}^{-1} - 1}{2}.$$

Solving for b_n^2 yields the Spearman–Brown expression (Spearman, 1910; Brown, 1910). For a unit probe, the decomposition $\hat{w}_{g,n} = b_n u_g + \zeta_{g,n}$ has $\mathbb{E}[\zeta_{g,n} \mid u_g] = 0$ and $\mathbb{E}\|\zeta_{g,n}\|^2 = 1 - b_n^2$. Thus $b_n^{-2} - 1$ is the ratio of residual variance to squared mean magnitude, and the stated reliability law specifies how this ratio scales with sample size. Replacing the consistencies by their sample estimates gives the attenuation-corrected estimator of λ^2.

For *(ii)*, the map $U_q \mapsto -U_q$ preserves the conditional law and exchanges the events $Y_q = +1$ and $Y_q = -1$. Therefore $P(Y_q = +1 \mid \mathcal{F}) = P(Y_q = -1 \mid \mathcal{F}) = 1/2$. Any deterministic prediction, or randomized prediction with independent randomness, has conditional accuracy $1/2$. For a predictive probability $p \in (0,1)$, conditional cross-entropy is

$$-\frac{1}{2}\log p - \frac{1}{2}\log(1-p) = -\frac{1}{2}\log[p(1-p)] \geq \ln 2,$$

with equality at $p = 1/2$. At $\lambda = 0$ in (M-u), the independent query residual is antipodally symmetric even after conditioning on the demonstrations and label-free query data, establishing the premise. □

Graph-conditional directions. If the direction depends on label-free graph covariates, $u_g = \lambda(O_g)w(O_g) + \sqrt{1-\lambda(O_g)^2}\,\xi_g$, the unconditional invariant mass is $\|\mathbb{E}[\lambda(O_g)w(O_g)]\|^2$ and can be small while every conditional direction is consistent across graphs. A small unconditional invariant mass therefore does not exclude transfer of a graph-conditional rule.

Similarity floor under graphwise symmetrization. For two graph-dependent rules, let $X_g = \langle u_t(g), u_e(g)\rangle$. Under the Gaussian feature model, Proposition 3 gives

$$a_g = \frac{1}{2} + \frac{\arcsin X_g}{\pi}, \qquad s = \frac{1}{2} + \frac{\mathbb{E}_g|\arcsin X_g|}{\pi}, \qquad s_{\text{global}} = \frac{1}{2} + \frac{|\mathbb{E}_g \arcsin X_g|}{\pi}.$$

The triangle inequality gives $s \geq s_{\text{global}}$. To compute the zero-signal floor, take orthogonal core directions $w_t \perp w_e$, equal invariant strengths λ, and independent residuals. Expanding the inner product gives

$$X_0 = \lambda\sqrt{1-\lambda^2}\big(\langle w_t, \xi_e\rangle + \langle \xi_t, w_e\rangle\big) + (1-\lambda^2)\langle \xi_t, \xi_e\rangle.$$

Each term has zero mean. In fact X_0 has a symmetric distribution. To see this, let R reflect across $w_t^{\perp}$. The transformation $(u_t, u_e) \mapsto (-Ru_t, Ru_e)$ preserves their independent direction laws, because $Rw_t = -w_t$ and $Rw_e = w_e$, and reverses their inner product. Since $\arcsin$ is odd, $\mathbb{E}\arcsin X_0 = 0$. Consequently

$$s_{\text{global}} = \frac{1}{2}, \qquad s_{\text{floor}} = \frac{1}{2} + \frac{\mathbb{E}|\arcsin X_0|}{\pi} \geq \frac{1}{2},$$

with strict inequality whenever $P(X_0 \neq 0) > 0$. This calculation locates zero global transfer signal on the graphwise similarity axis. A linear fit to gains over a finite similarity interval also depends on the curvature described in App. D.2.

Axis inflation affects fitted efficiency directly. If a predictor attains gain g_i but the fitted regressor is $f_i \geq g_i \geq 0$, then the slope through the origin satisfies $\hat{\eta} = \sum_i f_i g_i / \sum_i f_i^2 \leq 1$. In particular, monotonicity of the finite-context law gives $f_K(s_i) \geq f_K(s_{\text{global},i})$.

The Cora test. The Cora bank gives $r = +0.821$ and rejects the prediction based on its unconditional invariant mass, with the top quintile of measured gains averaging $\Delta_{\text{ICL}} = 0.239$. The screen therefore bounds the graph-invariant component of a task, and Cora transfers through a component that this screen does not measure. Amazon-ratings has a similarly small unconditional invariant mass and training loss at $\ln 2$, consistent with the symmetry condition of Proposition 4(ii). The excess of the fitted x intercept over the calculated similarity floor is $+0.005$ on Roman-empire (0.554 against 0.549), $+0.061$ on CiteSeer (0.645 against 0.584), and $+0.109$ on Cora (0.706 against 0.597), increasing as the estimated unconditional invariant mass decreases.

F THE EXTERNAL CHECKPOINT

Section 4.3 reports the frozen GraphPFN checkpoint at its pretraining scale. This appendix states the evaluation protocol (App. F.1) and reports the law at the default episode size and the sweep over episode sizes (App. F.2).

F.1 PROTOCOL

GraphPFN-1.3 (a LimiX backbone (Zhang et al., 2025) with graph adapters, about 20M parameters, Apache-2.0 license) runs frozen over exported episodes in its own pinned environment behind an episode/prediction file interface. Episode graphs are BFS subgraphs of the five real datasets (App. E.1). The node input is the leading eight principal components of the dataset's node features, standardized per graph and again over all nodes of the episode, and no structural descriptor is supplied. The context rules are the forty rules of the feature ladder (rule ids 124 to 163) and the evaluation tasks are the two held-out feature tasks of that ladder, so the similarity axis is the feature-ladder agreement measured on the query graphs. Support and query graphs enter as a disconnected union (Section 4.3), and labels enter the model only through its context interface. The context ratio, the share of union nodes that carry a label, is about 7% at K=16 on the default unions and is held constant in the sweep of App. F.2. It lies inside the checkpoint's pretraining range. The external-model evaluation uses a paired support-permutation baseline, $\Delta_{\text{ICL}} := \text{acc}(\text{intact}) - \text{acc}(\text{support-permuted})$ on the same episodes. Because the support is labeled by the context rule itself, polarity recovery involves no vote against a stored rule, and the efficiency is the asymptotic form $\eta = \Delta_{\text{ICL}}/(s - \frac{1}{2})$, which equals $2q_K - 1$ for a learner that induces the context rule with per-node fidelity q_K (Lemma 1 with $\gamma = 0$). The pretraining pool of a frozen checkpoint is fixed in this experiment, so the similarity axis is placed in the context. Support nodes

are relabeled by the graded rule, the query keeps its evaluation-rule label, and the exact query-graph agreement of each cell is stored at export.

F.2 The episode-size sweep

At the default episode size (m=4, unions of about 240 nodes, K=16, 600 episodes per cell) the law holds on four of the five datasets, with $r(s, \Delta_{\mathrm{ICL}}) = +0.85$ to $+0.94$ per dataset. Roman-empire gives $r = +0.45$ at this size and $r = +0.94$ or higher from m=8 on. The pooled 25-cell r is $+0.759$ with pooled permutation $p = 10^{-4}$, the pooled support-permutation accuracy is 0.497, and the efficiencies are $\eta = 0.25$ to 0.54 (Table 13, Figure 7a). These unions are smaller than the graphs used in pretraining. The sweep over episode sizes ($m = 4$ to 32 demonstration graphs, unions of 240 to 1,600 nodes, $K = 4m$ at constant context ratio, 200 episodes per cell) was pre-registered on Cora and PubMed and extended to the other three datasets at the same protocol. The efficiency rises with m on all five datasets, to 0.93, 0.97, 0.97, 0.96, and 0.88 on Cora, CiteSeer, PubMed, Roman-empire, and Amazon-ratings at m=32 (Figure 7b), and the law holds at every size from m=8 on, with $r = +0.82$ to $+0.997$ at m=32. Along this sweep K grows with m. A comparison at fixed $K = 32$ separates the number of labeled examples from the size of the union. With $m = 8$ graphs and four labels per graph the efficiencies are 0.77, 0.87, 0.84, 0.72, 0.71, and with m=32 graphs and one label per graph, which lowers the context ratio from 7% to 2%, they are 0.68, 0.68, 0.90, 0.62, 0.65, so enlarging the union at constant K lowers the efficiency on four datasets and raises it by 0.06 on PubMed, and the rise along the sweep is a rise with the number of labeled examples. Each per-level gain is a paired difference of accuracies estimated from 200 episodes.

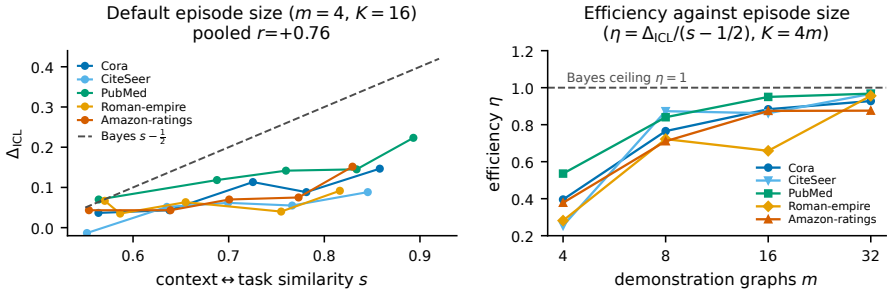

Figure 7: The frozen GraphPFN checkpoint at the default episode size and across episode sizes. (a) Δ_{ICL} against context–task similarity at m=4 on five real datasets (pooled $r = +0.759$), with the Bayes line. (b) Efficiency $\Delta_{\mathrm{ICL}}/(s - \frac{1}{2})$ against the number of demonstration graphs m for the five datasets; the sweep was pre-registered on Cora and PubMed and extended to the other three datasets.

G Implementation details

Architectures. Figure 8 illustrates the four in-context architectures. Model A uses a 4-layer GraphSAGE encoder (Hamilton et al., 2017) (hidden 64) over node features concatenated with standardized structural descriptors and a role embedding (support anchor or masked); a projection head; and a metric readout that classifies the query by distance to per-class support prototypes (Snell et al., 2017) (soft-NN and attention-pooling variants behave equivalently). Model B uses the same encoder with the readout replaced by a PRODIGY-style metagraph, in which per-episode label nodes are connected to support embeddings with signed edges (own class $+$, other class $-$, query edges 0), a two-layer, four-head attention GNN of width 64 with layer normalization runs over this task graph, label nodes are mean-centered, and the query is scored by cosine with a learned scale. Model C1 uses a set transformer (Lee et al., 2019) (two pre-norm blocks, four heads, d_{model}=64, FFN 128, dropout 0.1) over one query token and the support tokens with class-label embeddings. Its head forms a prototype per class as the masked mean of that class's support outputs, replaces the prototype of a class without support by a learned shared null vector, L2-normalizes the prototypes and the query output, and scores class c by $-\tau\|\hat{z}_q - \hat{p}_c\|^2$ with a learned temperature τ. At K=0 both prototypes equal the null vector and the logits tie exactly (Remark 1). Model C2 uses the same blocks (Vaswani et al., 2017) over all episode nodes with block-diagonal episode attention and a

Table 13: The law on the frozen GraphPFN-1.3 checkpoint. The top block gives the default episode size (m=4, K=16, five levels $\times$ 600 episodes per dataset). The bottom block gives the episode-size sweep (m demonstration graphs, K=$4m$, 200 episodes per cell). $\Delta_{\mathrm{ICL}} := \mathrm{acc} - \mathrm{acc}_{\mathrm{perm}}$; $\eta = \Delta_{\mathrm{ICL}}/(s-\frac{1}{2})$; exact permutation p over the 5! level orderings. At the default size the pooled 25-cell r is $+0.759$ with pooled mean-Fisher-z permutation $p = 10^{-4}$ and pooled support-permutation accuracy 0.497. At m=32 the pooled support-permutation accuracy is 0.499.

	Cora	CiteSeer	PubMed	Roman-emp.	Amazon-rat.
$r(s, \Delta_{\mathrm{ICL}})$	$+0.919$	$+0.889$	$+0.936$	$+0.445$	$+0.854$
$r(s, \mathrm{acc} - \frac{1}{2})$	$+0.933$	$+0.856$	$+0.921$	-0.111	$+0.771$
exact p	.025	.025	.008	.267	.017
η	0.396	0.254	0.536	0.282	0.378
episode-size sweep (m demonstration graphs, K=$4m$, 200 episodes per cell):					
η at m=8 ($\sim$430 nodes)	0.765	0.873	0.840	0.722	0.711
η at m=16 ($\sim$820 nodes)	0.884	0.862	0.950	0.659	0.875
η at m=32 ($\sim$1,600 nodes)	0.928	0.966	0.968	0.956	0.876
r at m=32	$+0.819$	$+0.972$	$+0.904$	$+0.962$	$+0.997$
exact p at m=32	.042	.008	.008	.033	.008
support-permutation acc. at m=32	0.488	0.507	0.504	0.488	0.508

graph-index embedding for up to 16 graphs, with the same head. In Model A and Model B the class labels enter only through the readout, either the prototypes or the metagraph, and the encoders receive a role embedding (support anchor or masked) and no class label. In Model C1 and Model C2 the class label of each support node enters as a class embedding added to the support token. The query label is masked in all four models.

Training. Adam, learning rate $3{\cdot}10^{-3}$, 15 epochs $\times$ 200 steps $\times$ 32 episodes per batch, K drawn per episode from $\{0, 1, 4, 8, 16\}$; binary classes; feature dimension 8; graph sizes $n \in [32, 64]$; m=4 demonstration graphs per episode; families ER/SBM/BA/WS unless noted. Support nodes are drawn per demonstration graph by a class round-robin that starts at a uniformly random class, with the picks shuffled. Seeds are fixed (seed 0 for the single-rule ladders and the width sweep, seeds 0 to 4 for the five-seed families, seeds 0 to 2 for the three-seed families), and every run directory stores its resolved Hydra configuration. Table 14 lists the evaluation protocol of every result family.

Table 14: Evaluation protocol per result family. "Episodes" is the number of evaluation episodes per support size and cell. The mixture controls train arm C for 8 epochs and arm R for 30 epochs, and the width-sweep control trains the width-256 Model C1 for 30 epochs. Mixture seeds are 5 for arms C and R and 3 for arm S and the controls. The real-label banks have 40 tasks on Roman-empire and 33, 34, and 28 on CiteSeer, Cora, and Amazon-ratings.

result family	models	training pool	seeds	episodes	section
single-rule ladders	A, B, C1, C2	T=1, 40 rules	1	600	2.2, C.2, D.1
seed replication	A	T=1, 40 rules	3	600	C.2
width sweep	A, C1	T=1, 40 rules	1	600	C.4
T=8 pool sweep	A	T=8, 5 pools	5	1,200	C.1
polarity switch	A	T 1 to 8	5	1,200	C.1
polarity $\times$ similarity	A	T 1 to 8	3	500	D.4
family diversity (M)	A, B, C1, C2	T=4, M 1 to 8	5	1,200	D.3
coverage-controlled T	A, B	T 1 to 16	5	1,200	D.3
mixture arms	A	T=3, S T=1	5 / 3	600	D.5
real structure	A	T=8, 5 pools	3	500	E.3
real features	A	T=8, 5 pools	3	1,200	E.3
real labels	A	T=1, one task per run	3	600	E.3, E.5
GraphPFN default / sweep	frozen	context rule, feature ladder	–	600 / 200	F

Control definitions. Zero-shot accuracy is measured at $K = 0$. Label propagation (Zhu & Ghahramani, 2002) iteratively propagates support labels over the union graph and returns the query's

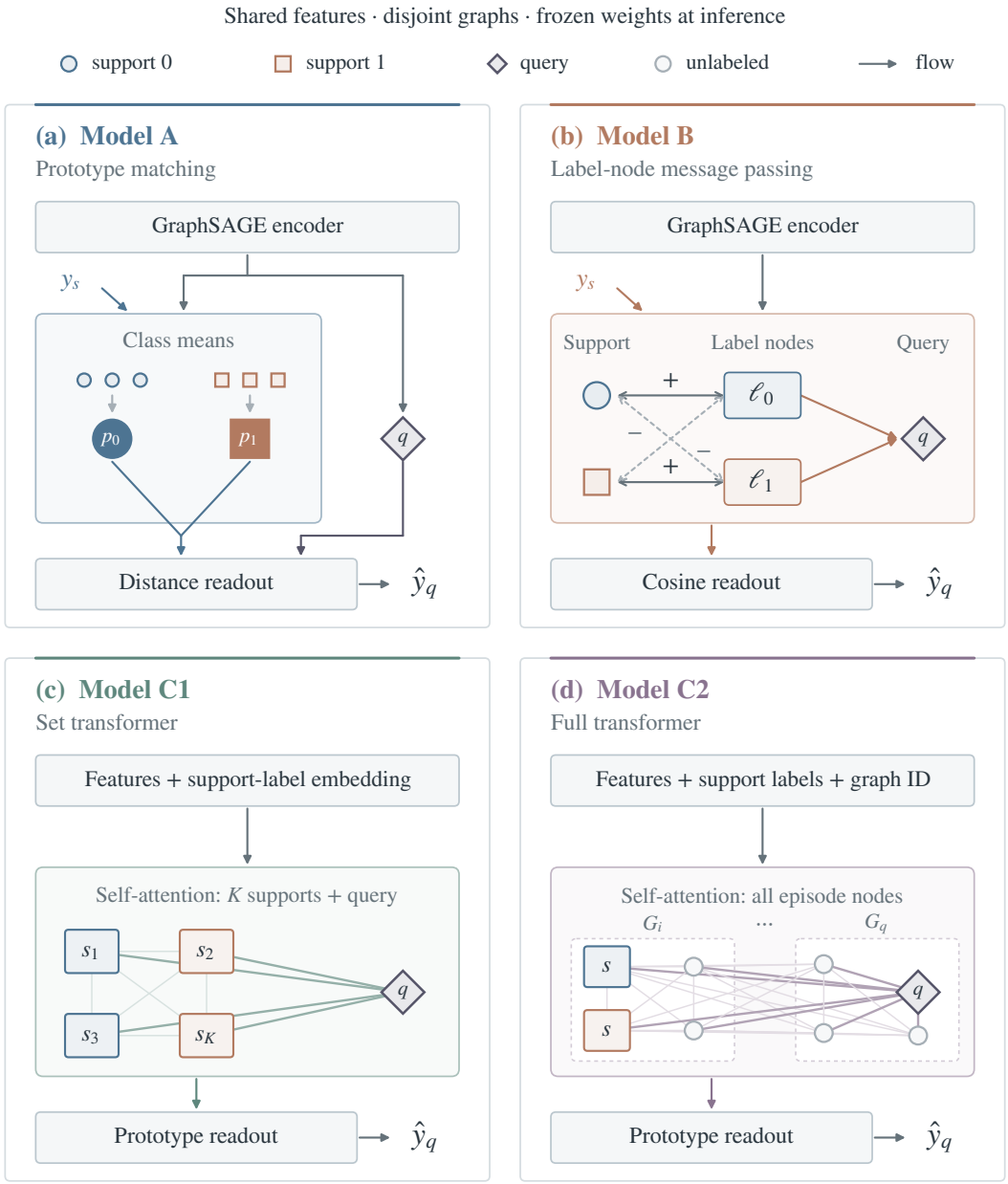

Figure 8: **The four model classes.** The panels show how each model conditions the query prediction on the support labels. (a) Model A aggregates graph-encoded support nodes into class prototypes and compares them with the query. (b) Model B uses the same encoder architecture with a signed metagraph that passes support information through label nodes to the query. (c) Model C1 applies self-attention to support and query tokens. (d) Model C2 attends to all episode nodes, with graph-index embeddings. Both transformers use a prototype readout. Connections inside the attention blocks represent attention between tokens from disjoint input graphs. Query labels are masked; $\hat{y}_q$ denotes the predicted query label.

argmax class. The support-permutation control permutes support labels within each episode before reevaluation. The wrong-rule control relabels supports with a different rule and keeps the query label. The feature-separability control uses cross-validated logistic regression from query features alone. The support-order control fixes the starting class of the class round-robin sampler at 0 and reevaluates on paired episodes.

Compute. All experiments were run on a single workstation with four NVIDIA RTX 3090 GPUs.