LABEL SHARING THROUGH SUFFICIENT GRAPH EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive substructures such as functional groups recur across graphs with different backgrounds, so graphs sharing an explanation could share labels. Existing guarantees require such graphs to carry nearly the same label. We show that label sharing needs only sufficiency, meaning that the rest of the graph adds no label information once the explanation is known. Our learner, pooled ERM, gives each training graph the label frequency of its explanation class, fits the resulting weighted majority targets, and averages fitted predictions within observed classes. Since pooling reuses the fitting sample, the proof separates within-class representatives from class counts. Let d be the VC dimension of a binary class $\mathcal{H}$ across the classes of an explanation map T. If $I(Y; G \mid T(G)) \leq \kappa$, the learner returns with probability $1-\delta$ a deterministic classifier within $3\sqrt{\kappa/2} + \epsilon$ of the best risk in $\mathcal{H}$ from $O((d + \log(1/\delta))/\epsilon^2)$ labels, even when the ordinary VC dimension is infinite. At $\kappa = 0$ label noise within classes is unrestricted, and the d/ϵ^2 dependence cannot be improved in general. Risk-dependent rates refine this bound, and a rate of order $d \log m/m$ holds under exact sufficiency, Bayes realizability, and a positive margin. With supplied explanations, GNNs trained by pooled ERM improve significantly in every controlled setting, with large gains under label noise. With mined keys on real graphs, pooling only within gated keys raises accuracy significantly on TOX21 and AIDS under a strict label budget and never lowers it significantly.

1 INTRODUCTION

Labels for molecular graphs come from laboratory measurements, such as the Ames tests behind mutagenicity data (Kazius et al., 2005), while the substructures that drive those labels recur across many molecules. A toxicophore is a functional group associated with mutagenicity, and molecules with very different backgrounds can contain the same one. Such structures come from domain knowledge, pattern mining, or subgraph explanation methods (Ying et al., 2019; Luo et al., 2020; Miao et al., 2022). When graphs share an explanation, one label informs the others, so fewer labels could suffice. Yet a toxicophore raises the probability that a molecule is mutagenic without fixing its label, so two molecules sharing it can differ in label. Can recurring explanations reduce the number of labels needed to learn a graph classifier when labels within an explanation disagree?

Existing guarantees for label reuse assume that such disagreement is absent or rare. Learning under transformation invariances uses an across-orbit VC dimension (Shao et al., 2022), where the invariance makes labels deterministic on each orbit. Its extension to subgraph explanations requires two graphs with the same explanation to disagree in label with probability at most $\epsilon/32$ (Zheng et al., 2024). If 80% of the molecules containing a toxicophore were mutagenic, two of them would disagree with probability 0.32, far above this threshold for every accuracy target. Both learners override ERM with the labels of training graphs that share the query's explanation, so under noise the copied label can be the atypical one, and both sample complexities carry logarithmic factors in d.

What label sharing needs is sufficiency. An explanation $T(G)$ is sufficient when the rest of the graph carries no further label information once $T(G)$ is known, that is, $I(Y; G \mid T(G)) = 0$, while the label probability within an explanation class can take any value. We allow $I(Y; G \mid T(G)) \leq \kappa$ and measure complexity by d, the largest number of graphs with distinct explanations that the hypothesis class $\mathcal{H}$ can shatter, which can be finite when the ordinary VC dimension is infinite. Our goal is to learn from a number of labels set by d for any label noise within classes, under two constraints. 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.

learner must compete with the best hypothesis in the original class $\mathcal{H}$, which may use background features that the explanation ignores, and discarded information should enter the guarantee as an additive cost, since supplied or mined explanations may be only approximately sufficient.

Three obstacles prevent a direct extension of the earlier learners. First, no single training label represents a noisy class, so any rule that copies labels inherits their noise. Second, averaging labels within classes makes each fitted target depend on the whole sample, and averaging predictions makes the output depend on it again. Uniform convergence over $\mathcal{H}$ cannot absorb this dependence, since $\mathrm{VC}(\mathcal{H})$ may be infinite even when d is small. Third, averaging within a class replaces a hypothesis by its class mean, which can remove the background information that the best hypothesis in $\mathcal{H}$ uses, and explanations mined from data can merge graphs whose label probabilities differ.

We address these obstacles with pooled ERM and its analysis. For label noise, each training graph receives the label frequency of its explanation class as a soft target, equivalently a majority label weighted by its margin, and the fitted hypothesis is averaged over training graphs sharing the query's explanation, while unseen explanations keep its prediction. A molecule with a toxicophore is thus trained toward the mutagen frequency among training molecules sharing its explanation rather than toward its own label. For sample dependence, the proof draws one representative per class independently of the class counts, so the pooled losses have VC dimension at most d even when a representative appears many times. For imperfect explanations, pooling a fixed hypothesis changes its risk by at most the average gap between graph-level and class-level label probabilities, which Pinsker's inequality bounds by $\sqrt{\kappa/2}$, and for finitely many classes a risk decomposition identifies which mined groups are worth pooling. We summarize the three contributions below.

- Pooled ERM comes within $3\sqrt{\kappa/2} + \epsilon$ of the best risk in $\mathcal{H}$ with probability $1 - \delta$ from $O((d + \log(1/\delta))/\epsilon^2)$ labeled graphs in total. At $\kappa = 0$ this holds for arbitrary label noise within classes. The d/ϵ^2 dependence has no logarithmic factor and cannot be improved in general.
- Risk-dependent rates hold without Bayes realizability, and a rate of order $d \log m/m$ holds under exact sufficiency, Bayes realizability, and a positive margin. For finitely many classes, the risk decomposition yields a gating rule for selective sharing with any fallback classifier.
- Exact-risk simulations track noise, complexity, and approximation as predicted. With supplied explanations, pooled ERM improves GNNs significantly in all 94 controlled settings, by up to 43 points under 20% label noise. With keys mined outside the training labels under a strict label budget, pooling only gated groups raises accuracy significantly in 24 of 64 settings and never lowers it significantly, and gated majority sharing keeps its TOX21 gains over tuned GNNs.

2 Explanation Classes and Noisy Sufficiency

Let $(G, Y) \sim P$, with $Y \in \{0, 1\}$, and let $\mathcal{H}$ be a nonempty class of measurable binary graph classifiers. Its comparator risk is $R^*_{\mathcal{H}} = \inf_{h \in \mathcal{H}} \mathbb{P}\{h(G) \neq Y\}$. A probabilistic predictor $q(G) \in [0, 1]$ predicts label one with probability $q(G)$; its risk averages this prediction randomness:

$$R(q) = \mathbb{E}\ell(q(G), Y), \qquad \ell(q, y) = y + (1 - 2y)q. \tag{1}$$

For binary q, this is ordinary classification error.

Structural keys. A fixed measurable map T assigns each graph to an explanation key. It may encode a supplied subgraph or the joint occurrence of selected patterns S:

$$T_S(G) = \{p \in S : p \text{ occurs in } G\}. \tag{2}$$

Equality of keys defines disjoint classes, including the empty key. For supplied subgraphs $\Psi(G)$, viewed up to attributed isomorphism, containment realizes equality of explanation classes when

$$\Psi(G) \subseteq G' \implies \Psi(G') = \Psi(G) \quad \text{on the graph support.} \tag{3}$$

This compatibility condition (Zheng et al., 2024) involves the map, support, and matching rule.

Sufficiency. We quantify discarded label information (Zheng et al., 2024) by

$$I(Y; G \mid T(G)) \leq \kappa. \tag{4}$$

Information is in nats. At $\kappa = 0$, Y and G are conditionally independent given $T(G)$; the label probability $\eta_t = \mathbb{P}(Y = 1 \mid T = t)$ can still take any value in $[0, 1]$. Thus intrinsic label noise is

unrestricted. The main results assume a fixed map. They also hold conditionally for a map learned on independent earlier data, whose labeling cost is additional. Appendix A.4 permits selection from a fixed countable family, paying a prior or description-length cost for the selected map.

Across-class complexity. Let $d = \mathrm{VC}_T(\mathcal{H})$ be the largest size of a set shattered by $\mathcal{H}$ whose keys are pairwise distinct. This is the across-orbit VC measure (Shao et al., 2022) expressed for a partition, and the explanation-assisted VC dimension (Zheng et al., 2024). It satisfies

$$d \le \mathrm{VC}(\mathcal{H}), \qquad d \le J \quad \text{if } T \text{ has } J \text{ values.} \tag{5}$$

Within-class shattering can be arbitrarily large, so d can be far below $\mathrm{VC}(\mathcal{H})$. We assume standard Borel input and key spaces, and measurable learning rules and class suprema throughout.

3 Learning by Pooling within Explanation Classes

3.1 Pooling and Fitting on the Same Sample

Let $D = \{(G_i, Y_i)\}_{i=1}^m$ be an i.i.d. labeled sample. Write $N_t = |\{i : T(G_i) = t\}|$ and $\bar{Y}_t = N_t^{-1} \sum_{i:T(G_i)=t} Y_i$ for a represented class. Define

$$Q_D h(G) = \begin{cases} N_{T(G)}^{-1} \sum_{i:T(G_i)=T(G)} h(G_i), & N_{T(G)} > 0, \\ h(G), & N_{T(G)} = 0. \end{cases} \tag{6}$$

Pooled ERM returns $\hat{q} = Q_D \hat{h}$, where

$$\widehat{R}_D(Q_D \hat{h}) \le \inf_{h \in \mathcal{H}} \widehat{R}_D(Q_D h) + \xi \tag{7}$$

for an optimization tolerance $\xi \ge 0$. The fitting objective has the exact identity

$$\widehat{R}_D(Q_D h) = \frac{1}{m} \sum_{i=1}^m \ell(h(G_i), \bar{Y}_{T(G_i)}). \tag{8}$$

The same objective averages empirical loss over all within-class label permutations. Equivalently, let $c_t = \mathbf{1}\{\bar{Y}_t > 1/2\}$ and $w_t = |2\bar{Y}_t - 1|$. Up to a term independent of h, the objective is $m^{-1} \sum_i w_{T(G_i)} \mathbf{1}\{h(G_i) \ne c_{T(G_i)}\}$. Tied groups have weight zero. This is ordinary weighted ERM.

Theorem 3.1 (Learning from noisy sufficient explanations)**.** *For any distribution P, fixed T, and nonempty class $\mathcal{H}$ with $d < \infty$, pooled ERM satisfies, with probability at least $1 - \delta$,*

$$R(\hat{q}) \le R^*_{\mathcal{H}} + 3\sqrt{\kappa/2} + \xi + C\sqrt{\frac{d + \log(1/\delta)}{m}}, \tag{9}$$

where C is universal and $0 < \delta < 1/2$. With exact ERM and fresh validation for threshold selection, it returns a deterministic classifier $\hat{g}$ satisfying

$$\mathbb{P}\{R(\hat{g}) \le R^*_{\mathcal{H}} + 3\sqrt{\kappa/2} + \epsilon\} \ge 1 - \delta \tag{10}$$

using $O((d + \log(1/\delta))/\epsilon^2)$ labeled graphs in total.

At $\kappa = 0$, graphs within a class need only share a label distribution, so intrinsic noise is unrestricted. The d/ϵ^2 dependence then cannot be improved in general, since an injective map gives $\kappa = 0$ and $d = \mathrm{VC}(\mathcal{H})$, where classical agnostic lower bounds apply (Devroye et al., 1996), and it remains necessary with deterministic labels (Shao et al., 2022). The pooled and thresholded outputs need not belong to $\mathcal{H}$, and fitting reduces to weighted ERM over $\mathcal{H}$ on group-majority targets.

Controlling sample reuse. Write $\eta_G = \mathbb{P}(Y = 1 \mid G)$ and define the auxiliary risk $R_0(q) = \mathbb{E}\ell(q(G), \eta_{T(G)})$. The difficult term is the feature deviation between $R_0(Q_D h)$ and its empirical counterpart, uniformly over h. Introduce an independent ghost feature sample and condition on their pooled $2m$ graphs. A class containing k_t fitting and l_t ghost graphs contributes

$$\mathbb{E}_r\big[(l_t - k_t)\ell(h(r_t), \eta_t)\big]/m, \tag{11}$$

where r_t is uniform among its fitting graphs, or among its ghost graphs if $k_t = 0$. Conditional on the complete count vector, these representatives are independent uniform draws from the pooled classes. Their law does not depend on the count vector.

We can therefore draw the representatives first and assign sample roles afterward. A random pairing and independent signs generate a uniform balanced assignment. Every class then has only one distinct representative, so the relevant loss vectors have VC dimension at most d, even when that representative appears many times. The VC entropy bound and bounded differences give an exponential deviation of order $\sqrt{(d + \log(1/\delta))/m}$ (Haussler, 1995; Bartlett & Mendelson, 2002). Conditional on the fitting features, independent label noise admits the same complexity control. Equation (11) includes classes absent from fitting, whose query predictions retain h.

Preserving the comparator. For a fixed h, the difference between its mean pooled training loss and $R(h)$ is a weighted sum of conditional covariances between $h(G)$ and η_G. It is at most $\Delta_T = \mathbb{E}|\eta_G - \eta_{T(G)}|$. Replacing one example changes that empirical loss by at most $3/m$, giving concentration for this one comparator without controlling the unpooled class uniformly. The two conditional-mean comparisons add at most $2\Delta_T$; conditional Pinsker gives $\Delta_T \le \sqrt{\kappa/2}$. These steps prove (9).

For deterministic output, $g_u = \mathbf{1}\{\hat{q} \ge u\}$ satisfies $R(\hat{q}) = \int_0^1 R(g_u)\, du$. Fresh validation selects a threshold whose error is close to the best threshold's error, using $O(\log(1/\delta)/\epsilon^2)$ additional labels. Appendix A gives explicit constants, the dependence argument, and complete sample accounting.

3.2 Risk-Dependent and Fast Rates

Pooling also admits a risk-dependent refinement under approximate sufficiency. Let $\Lambda = (d + 1)\log(2em) + \log(1/\delta)$. With m fitting and m fresh validation labels, exact pooled ERM followed by threshold selection satisfies

$$R(\widehat{g}) \le R^*_{\mathcal{H}} + 3\Delta_T + C\left(\sqrt{\frac{(R^*_{\mathcal{H}} + \Delta_T)\Lambda}{m}} + \frac{\Lambda}{m}\right) \tag{12}$$

with probability at least $1 - \delta$. The proof controls squared deletion changes of a fixed comparator's pooled loss by that loss itself. Combined with feature and label offsets, this gives an estimation cost that decreases with $R^*_{\mathcal{H}} + \Delta_T$, without Bayes realizability or knowledge of $R^*_{\mathcal{H}}$ and Δ_T.

A fast excess-Bayes-risk guarantee also holds when the Bayes decision is represented in $\mathcal{H}$ and conditional label probabilities stay away from one half. This is the Massart condition (Massart & Nédélec, 2006); the complexity remains d.

Theorem 3.2 (Fast learning with sufficient explanations)**.** *Assume $Y \perp G \mid T$, a Bayes classifier $h^* \in \mathcal{H}$, and $|2\eta_{T(G)} - 1| \ge \gamma > 0$ almost surely. For the same pooled ERM, let $\hat{g} = \mathbf{1}\{Q_D\hat{h} \ge 1/2\}$. With probability at least $1 - \delta$,*

$$R(\hat{g}) - R^* \le 8\xi + \frac{C\{d\log(2em) + \log(2/\delta)\}}{\gamma m}, \tag{13}$$

where $R^* = R(h^*)$ and C is universal. The learner needs neither γ nor validation labels.

The proof uses the same representatives with an offset: subtracting a multiple of empirical excess loss makes the exponential moment of each fixed disagreement pattern at most one. There are at most $\sum_{j=0}^{\min(d,2m)} \binom{2m}{j}$ such patterns. Feature and label offsets then combine through the ERM inequality. Appendix A.1 gives the explicit constants.

With exact ERM, exact sufficiency, and Bayes realizability, a polynomial margin tail $\mathbb{P}(|2\eta_T - 1| \le u) \le Au^\alpha$ for $0 < u \le 1$ gives excess risk $O((\Lambda/m)^{(\alpha+1)/(\alpha+2)})$ for fixed $A \ge 1, \alpha > 0$, without margin tuning (Appendix A.1).

A decision-sufficiency extension permits unequal posteriors within each key: if a Bayes rule in $\mathcal{H}$ is keywise constant and $|2\mathbb{P}(Y = 1 \mid G) - 1| \ge \gamma$, Appendix A.1 gives a bound with γ^{-2} dependence. The general and risk-dependent oracle bounds continue to allow a misspecified $\mathcal{H}$.

3.3 Selective Majority Sharing with a Fallback

For finitely many keys, fitting a separate score to each group's shared labels selects its empirical majority. A practical learner can use these votes on selected groups and retain a trained graph predictor b elsewhere. Let A be any data-dependent subset of observed keys with a unique majority, and let h_A use that majority on A and b outside it.

Write $R_A(f) = \mathbb{E}[\mathbf{1}\{T(G) \in A\}\ell(f(G), Y)]$ for risk contributed by the selected region. Let g^* be a Bayes rule using the full graph and c_T^* a Bayes rule using only its key. Define

$$V_A(b) = R_A(b) - R_A(g^*), \qquad \Gamma_T(A) = R_A(c_T^*) - R_A(g^*).$$

Thus $V_A(b)$ is the baseline error available to remove, above irreducible noise, and $\Gamma_T(A)$ is the classification cost of using one label per key. Both include the selected region's probability mass. The second vanishes when a Bayes decision is constant within each selected class.

Proposition 3.3 (Selective label reuse)**.** *For fixed T taking $J < \infty$ values almost surely and m i.i.d. fitting examples, any such gate and any data-dependent baseline satisfy*

$$\mathbb{E}R(h_A) \le \mathbb{E}R(b) - \mathbb{E}V_A(b) + \mathbb{E}\Gamma_T(A) + E_m(T), \tag{14}$$

where, for $p_t = \mathbb{P}(T = t)$ and $\gamma_t = |2\eta_t - 1|$,

$$E_m(T) = \sum_t p_t \gamma_t e^{-m p_t \gamma_t^2/4} \le \sqrt{J/m}. \tag{15}$$

If every $\gamma_t \ge \gamma > 0$, then $E_m(T) \le 4J/(em\gamma)$. If labels are deterministic given T, sharing on observed keys never increases the baseline's risk, for every fitting set. Sharing on all observed keys then gives

$$R(h_A) \le U_D := \sum_t p_t \mathbf{1}\{N_t = 0\}, \qquad \mathbb{E}U_D = \sum_t p_t(1 - p_t)^m. \tag{16}$$

No independence of the gate or baseline from fitting data is needed. Write $K_A = \sum_{t \in A} p_t I(Y; G \mid T = t)$. Then $\Gamma_T(A) \le \sqrt{\mathbb{P}(T \in A) K_A/2}$: only information discarded inside selected groups contributes. Thus selective reuse can help without global sufficiency. Appendix A.3 proves the exact decomposition and the majority bound by Hoeffding and binomial averaging. The finite-class and missing-mass calculations are classical (Devroye et al., 1996; McAllester & Ortiz, 2003).

For uniform classes, expected uncovered mass is $(1 - 1/J)^m$, independent of the number of backgrounds. Key maps on real graphs take finitely many values, so the graph experiments test this regime, while the numerical study isolates classes whose dimension d stays small as their number grows.

4 Structural Label Sharing on Graphs

When explanations are unavailable, the implementation mines structural keys and uses their majority votes with a trained GNN. Proposition 3.3 analyzes this selective sharing mechanism for a fixed partition; pattern discovery supplies the partition in practice.

Selecting groups. The candidate vocabulary contains rings, attributed paths, or Weisfeiler–Leman descriptors (Shervashidze et al., 2011). For a selected pattern set S, the joint key (2) records the presence and absence of every selected pattern. Let $n_{t,c}$ be the labeled count in group t, $n_t = n_{t,0} + n_{t,1}$, $n_{\text{sel}} = \sum_t n_t$, and $\widehat{q}_t = (\max_c n_{t,c} + 1)/(n_t + 2)$. Selection greedily adds patterns that improve

$$F(S) = \sum_{t \in \mathcal{E}} \frac{n_t}{n_{\text{sel}}} (\widehat{q}_t - a), \tag{17}$$

where $\mathcal{E}$ contains groups meeting the support, unique-majority, and purity requirements. In the model-referenced configuration, a is the GNN validation accuracy. Pattern and group caps limit selection. Fitting labels determine the final votes; an empty selected vocabulary retains the baseline. Appendix C.5 specifies the protocol.

Combining predictions. For eligible voting keys A, stored majority labels c_t, and the same trained GNN b, the hybrid is

$$h_A(G) = \begin{cases} c_{T(G)}, & T(G) \in A, \\ b(G), & \text{otherwise.} \end{cases} \tag{18}$$

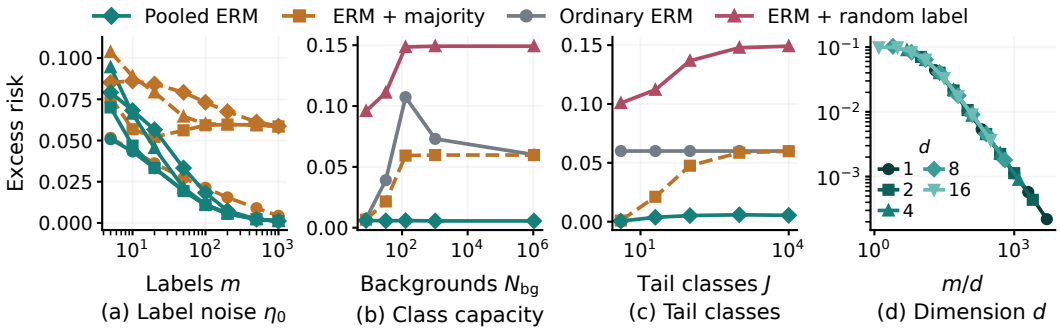


Figure 1: Excess risk of pooled ERM and three baselines on the same labeled sample. (a) Growing sample under recurrent-class noise $\eta_0 = 0.05$, 0.20, 0.35, and 0.45, marked by circles, squares, triangles, and diamonds. (b,c) Growing background capacity and number of tail classes at $m = 200$, with $d = 1$. (d) Pooled ERM for d from 1 to 16, against labels per dimension m/d. Population risks are exact; bands and bars show two Monte Carlo standard errors of the mean.

The gate seeks groups whose shared label improves on the baseline while covering enough queries to matter. Proposition 3.3 expresses this as captured baseline excess, classification bias from grouping, and majority estimation error. Support and purity are empirical proxies for these terms.

Proposition 3.3 lets the gate and baseline depend on fitting data but takes the key map as fixed. Mining the vocabulary outside the training labels meets this condition and keeps the gains.

5 Experiments

We test how pooling uses repeated explanation classes and how its sample requirement depends on across-class structure. Three learners appear. Pooled ERM is the learner of Theorems 3.1 and 3.2; the numerical study solves it exactly, and on graphs it trains a GNN on objective (8). Selective pooled ERM pools only the groups that pass the gate and leaves every other graph ungrouped, following the decomposition of Proposition 3.3. Majority sharing with a GNN fallback is the hybrid (18), which Proposition 3.3 bounds for any gate. Table 1 evaluates pooled ERM, Table 3 its selective version, and Tables 2, 4, and 5 majority sharing, with the mined gate except on MUTAGENICITY.

5.1 Noisy Learning with Fixed Across-Class Complexity

We construct a finite graph encoding with one recurrent explanation class and J ordered tail classes. The recurrent class has N_{bg} possible backgrounds and mass 0.85; the tail classes share the remaining mass. Tail Bayes labels follow a threshold on the class index and are independently flipped with probability 0.1. In the recurrent class, the probability of label one is $\eta_0 \in \{0.05, 0.20, 0.35, 0.45\}$. Thus $\kappa = 0$ while intrinsic noise can be large.

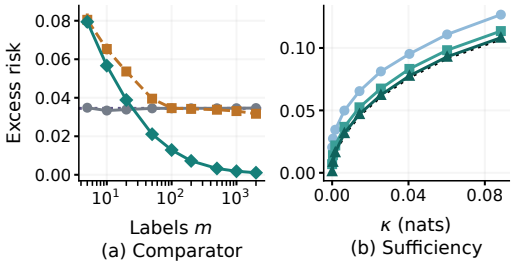


Figure 2: Comparator preservation and approximate sufficiency, with method colors as in Figure 1. (a) Excess over a non-invariant best hypothesis; the dotted line is the best invariant member of $\mathcal{H}$. (b) Excess of pooled ERM as the key loses label information, for $m = 50$, 200, and 1000 from light to dark; the dotted line is the projection cost. Bars show two Monte Carlo standard errors.

The hypothesis class contains every tail threshold with recurrent-class prediction zero, and arbitrary background classifiers whose tail prediction is one. Its across-class VC dimension is $d = 1$ for every N_{bg}, J, while ordinary VC dimension is N_{bg}. The best hypothesis is the true tail threshold. We solve ordinary and pooled ERM exactly on the same m labeled graphs. Pooled ERM uses these graphs for both aggregation and fitting. Comparisons apply majority or random-label sharing to ordinary ERM; the former extends the deterministic-label ERM-INV proce-

Table 1: Pooled ERM with supplied keys under 20% label noise. Entries are accuracy gains in points over the GNN trained on the same labels, and every gain is significant at the 5% level (paired sign-flip test over 30 runs, Holm-corrected). The last two columns give pooled ERM minus test-time voting at $t_n = 20$; an asterisk marks a difference that is also significant at the same 5% level.

		GCN, t_n			GIN, t_n			vs. voting	
Keys	J	2	10	50	2	10	50	GCN	GIN
FOUR-MOTIF	16	+9.67	+24.94	+30.32	+7.78	+23.36	+34.32	+2.18	+3.63*
$L = 2$	4	+19.58	+41.14	+35.23	+19.68	+40.06	+40.82	+1.66	+1.82
$L = 6$	22	+8.08	+21.09	+35.54	+6.66	+18.38	+37.44	+1.37	+3.25*
$L = 10$	56	+5.81	+16.48	+29.30	+4.36	+13.64	+27.41	+2.00	+2.61

Table 2: Accuracy gain in points of the hybrid over the GNN under a strict total label budget B. Both use the same B labels, and the GNN trains on every label outside the checkpoint set. An asterisk marks a gain significant at the 5% level in a paired sign-flip test over 30 runs, Holm-corrected across the 16 settings of each backbone. DHFR at $B = 1000$ uses its full pool of 681 labels. Appendix D.4 gives accuracies, coverage, balanced accuracy, and paired win and loss counts.

	AIDS		DHFR		TOX21		PROTEINS	
B	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN
100	+0.95*	+0.65	−0.13	+0.98	+3.00*	+2.42*	+1.22	+0.45
200	+2.83*	+1.93*	+1.29*	+1.73*	+3.74*	+2.16*	+1.07	+0.71
500	+3.45*	+0.78*	+1.29*	+0.49	+3.04*	+1.25*	+1.37*	+0.51
1000	+2.27*	+0.20	+0.62	+0.09	+2.96*	+0.25	+0.42	+0.09

dure (Shao et al., 2022) to group majorities. All methods see the same sample. Risks are exact, so error bars reflect only sampling variation.

Figure 1 shows pooled fitting approaching the best hypothesis at every noise level. Increasing background capacity or the number of tail explanations preserves its ability to learn the across-class threshold. This tests the role of d when the nominal group count is much larger. A product construction varies d from 1 to 16; its curves nearly coincide when plotted against m/d in panel (d). Appendix B gives the encoding, class, exact optimization, 32 noise configurations with 3,000 repetitions each, and 40 capacity configurations with 2,000 repetitions each. Larger-budget studies vary sample size, a uniform conditional margin, and the mass of nearly ambiguous keys. Their curves are consistent with the fast-rate and margin-tail bounds; Appendix B.2 gives all 118 configurations.

A second construction makes the best hypothesis depend on recurrent-class background features. Figure 2(a) shows pooled ERM approaching its risk, below the best invariant member of $\mathcal{H}$. Panel (b) varies within-key label probabilities; the large-sample excess approaches the comparator's projection cost. Appendix B.1 gives both grids. A further misspecified construction varies $R^*_{\mathcal{H}}$; its upper excess is consistent with the $\sqrt{R^*_{\mathcal{H}}/m}$ scale in the risk-dependent bound (Appendix B.4).

5.2 Supplied Explanations on Graphs

Protocol. We evaluate GCN, GIN, GraphSAGE, and PNA (Kipf & Welling, 2017; Xu et al., 2019; Hamilton et al., 2017; Corso et al., 2020) on three stratified 80/10/10 splits with ten seeds each. Eight nested fitting budgets range from $t_n = 2$ per class to the full pool. Pooled ERM trains the same network from the same initialization on the weighted objective (8) and averages its predictions within observed keys. Every method selects checkpoints with validation labels, and the hybrid also selects its candidate cut and pattern cap from 36 configurations on these labels and votes with fitting labels, so all access $2t_n + |D_{\text{val}}|$ distinct labels. Appendix C details the splits, budgets, and selection rules.

Compatible matching. On BA2MOTIFS, supplied motif labels are transferred using ordinary or induced subgraph matching, with support and purity gates disabled. At two fitting graphs per class, induced matching reaches accuracy 1.000 ± 0.000 for both GIN and GCN, over ten seeds on the original split. The corresponding GNN accuracies are 0.636 ± 0.155 and 0.500 ± 0.015; ordinary

Table 3: Selective pooled ERM with mined keys under the strict budget B. Entries are accuracy gains in points over the GNN trained on the same labels, and SAGE denotes GraphSAGE. An asterisk marks a gain significant at the 5% level in a paired sign-flip test over 30 runs, Holm-corrected within each pre-registered family. Appendix E.2 gives accuracies, key statistics, and the hybrids.

	AIDS				DHFR			
B	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE
100	$+1.88^*$	$+0.90$	$+3.17^*$	$+3.17^*$	$+0.76$	$+0.89$	-0.93	$+1.64$
200	$+3.83^*$	$+2.90^*$	$+5.05^*$	$+4.98^*$	$+0.04$	$+1.07$	-0.00	$+0.67$
500	$+4.40^*$	$+1.48$	$+2.95^*$	$+4.92^*$	$+1.38$	-0.00	-0.80	$+1.16$
1000	$+2.65^*$	$+0.72$	-0.17	$+2.23^*$	$+0.93$	$+1.07$	-0.13	$+3.29^*$
	TOX21				PROTEINS			
B	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE
100	$+2.61^*$	$+2.40^*$	$+2.02$	$+2.56^*$	$+3.18^*$	$+0.36$	$+1.64$	$+1.85$
200	$+2.37^*$	$+2.33^*$	$+2.40^*$	$+1.54$	$+1.85^*$	$+0.09$	$+1.58$	$+1.19$
500	$+2.25^*$	$+1.72^*$	$+0.00$	$+0.81$	$+1.13$	$+1.28$	$+0.62$	$+0.57$
1000	$+1.65^*$	$+0.51$	$+0.61$	-0.25	$+1.01$	-0.36	-1.07	$+0.68$

matching gives 0.792 ± 0.113 and 0.671 ± 0.111. Induced matching distinguishes the supplied cycle and house motifs while retaining cross-background reuse. Appendix D.1 gives full results.

Pooled ERM with supplied keys. FOUR-MOTIF has 16 keys, its planted motif sets, and a controlled family varies the number of motif types $L \in \{2, 6, 10\}$, giving $J = 1 + L + \binom{L}{2}$ keys with graph size and node-type counts controlled. Fitting and validation labels are flipped with probability 0.2 and test labels are clean. Table 1 shows that pooled ERM improves on the GNN trained on the same noisy labels everywhere, with gains shrinking as J grows at small budgets, as the class-count dependence predicts. On FOUR-MOTIF it also exceeds test-time voting on the same keys for all four backbones at every budget up to 100 graphs per class, by 0.6 to 5.0 points. Without noise, all 32 FOUR-MOTIF settings of four backbones and eight budgets improve significantly, and pooling and voting nearly coincide (Appendix E).

Table 4: Majority sharing on MUTAGENICITY with supplied (S) or permuted (P) descriptor sets, in points over the GNN, averaged over ten seeds.

	S−GNN		P−GNN	
t_n	GCN	GIN	GCN	GIN
2	$+5.22$	$+4.23$	-4.62	-4.64
5	$+7.23$	$+5.15$	-5.47	-6.54
10	$+2.77$	$+2.54$	-9.45	-9.56
20	$+3.60$	$+2.08$	-13.37	-15.27
50	$+3.05$	$+1.73$	-15.68	-16.61
100	$+1.06$	$+2.12$	-18.59	-18.01

Toxicophore descriptors. On MUTAGENICITY, ten fixed descriptors approximate known toxicophores (Kazius et al., 2005). Graphs with the same descriptor set share their fitting-label majority, and the GNN decides absent or tied sets. Table 4 shows gains of 1.1 to 7.2 points at every budget up to 100 graphs per class. Permuting descriptor sets across graphs keeps group sizes but breaks their link to structure, and the same rule then loses 4.6 to 18.6 points. The supplied sets beat the permuted ones in all ten seeds of every cell, so the gain comes from the descriptors themselves.

5.3 Label Reuse under a Total Annotation Budget

Strict label accounting. The preceding fitting-budget protocol shares validation access between methods. We also report an experiment with one budget B for every label used in training, checkpoint selection, pattern selection, and voting. From the merged fitting and validation pool, B labels are drawn with the pool's class proportions. A stratified set of $v = \min(50, \lceil 0.2B \rceil)$ labels is reserved for checkpointing; the GNN trains on all $B - v$ remaining labels. The hybrid uses that same GNN, selects patterns with checkpoint labels, and votes with training labels. It reads no additional labels.

Table 2 reports all 32 task–backbone–budget cells. The hybrid improves mean accuracy in 31 of them, significantly in 17 after Holm correction, and reduces it significantly in none. Seven of the

eight TOX21 cells are significant, with gains of 1.25 to 3.74 points through $B = 500$, even though the baseline trains on every non-checkpoint label. TOX21 is the balanced AhR subset in Appendix C.

Pooled ERM with mined keys. We rebuild the key map from graphs outside the training labels, mining unlabeled pool and checkpoint graphs and selecting patterns with checkpoint labels, so the map is fixed given the training sample. These keys form two to ten groups of 27 to 383 training graphs on average, so we pool only the groups that pass the gate and, as in Proposition 3.3, only information discarded inside those groups matters. Table 3 shows that selective pooled ERM raises accuracy significantly in 24 of the 64 settings, nine of sixteen on TOX21 and twelve of sixteen on AIDS, and lowers it significantly in none. Voting with the fixed vocabulary stays within 2 points of the deployed hybrid in every setting, so mining on the training graphs is not needed for these gains.

Tuned networks and ablations. Table 5 replaces the recipe network by one tuned on validation labels. On TOX21 the hybrid keeps significant gains of 1.3 to 4.6 points at $t_n = 50$ and 0.9 to 3.2 points under the strict budget, while on AIDS the tuned network already reaches 94 to 99% and leaves few errors to share. Ablations at $t_n = 50$ over six tasks and four backbones support both parts of the rule. Without a gate, a predictor on selected patterns beats one on random patterns by 16.4 points on average, and the gate adds 11.5 points to group voting. Appendices D.6 and D.3 give the complete tuning study, budget curves, ablations, and controls. The rule also attaches to a WL-kernel SVM, and GSAT and DIR are compared in Appendices D.5 and D.8.

Table 5: The hybrid over a tuned GNN on TOX21 at $t_n = 50$. Tuned is accuracy in percent and Gain is in points, each significant after Holm correction.

Model	Tuned	Gain
GCN	61.46	+4.61
GIN	64.91	+3.11
PNA	69.54	+1.32
SAGE	63.91	+4.18

6 Related Work

Learning with invariances. The across-orbit VC measure comes from ERM-INV, which fits ordinary ERM and overrides predictions on observed orbits; its agnostic guarantee allows a misspecified class but assumes deterministic invariant labels (Shao et al., 2022). Explanation-assisted ERM brings this measure to graph explanations under condition (3) and needs two graphs with the same explanation to disagree with probability at most $\epsilon/32$, using $O(d \log^2 d/\epsilon^2 + \log(1/\delta)/\epsilon^2)$ labels (Zheng et al., 2024). Our bound allows any label noise within explanation classes, charges $3\Delta_T$ for discarded information, and is optimal in d and ϵ at $\kappa = 0$. Sufficient representations also appear in probabilistic symmetry (Bloem-Reddy & Teh, 2020), and group-averaged training with probabilistic label invariance appears in a coreset analysis under quotient coverage and Lipschitz assumptions (Ayres et al., 2026).

Averaging and finite-class learning. Orbit-averaged losses have been analyzed under a specified transformation distribution (Chen et al., 2020), learnability of classes obtained by averaging random functions has been studied (Anderson & Benedikt, 2026), and compression adapts across families of partial concept classes, comparing against partial hypotheses in a clustering application (Pour & Ben-David, 2026). Here the averaging law is estimated from the fitting sample, representatives independent of class counts control this dependence, and the comparator remains the original binary class. Margin-based rates (Massart & Nédélec, 2006; Tsybakov, 2004) and finite-class majority learning with missing-mass analysis (Devroye et al., 1996; McAllester & Ortiz, 2003) are classical; our offset argument controls the empirical pooling operator through across-class complexity.

Graph explanations. GNNExplainer, PGExplainer, and GSAT identify predictive graph substructures (Ying et al., 2019; Luo et al., 2020; Miao et al., 2022); DIR learns invariant rationales (Wu et al., 2022). We use supplied or mined structures to share labels with a trained graph predictor.

7 Conclusion

Label sharing across explanations needs sufficiency, not deterministic labels. Pooling supervision within explanation classes then learns at a rate set by across-class complexity for any label noise inside the classes, pays separately for the information an explanation discards, and admits sharper rates under further conditions. With supplied graph explanations, GNNs trained this way gain substantially, and with coarse mined keys, pooling within gated groups improves GNNs under a strict label budget, with a gain set by the baseline errors those groups cover, as Proposition 3.3 describes.

AI USE STATEMENT

OpenAI Codex assisted with theoretical development, proof writing and checking, literature retrieval, numerical experiment design, implementation and simulated data generation, figure preparation, manuscript revision, and layout verification. Claude (Anthropic), used through Claude Code, assisted with revision planning, the design of the added graph experiments, rewriting and editing of the manuscript, table and figure formatting, consistency checks between the text and the reported results, and simulated reviewing. The authors are responsible for the mathematical claims, empirical evidence, citations, and final manuscript.

REPRODUCIBILITY STATEMENT

Appendix A gives complete proofs, including all independence assumptions. The source package includes code, configurations, and exported results for the numerical studies described in Appendix B. It also includes the graph benchmark tables and curves, whose datasets, training, selection, and evaluation protocols are specified in Appendices C–D.

REFERENCES

Aaron Anderson and Michael Benedikt. From learnable objects to learnable random objects. *Journal of Machine Learning Research*, 27(200):1–55, 2026. URL https://www.jmlr.org/papers/v27/25-1196.html.

Luciano C. Ayres, José C. M. Bermudez, Sérgio J. M. de Almeida, and Ricardo A. Borsoi. Group-invariant coresets for data-efficient active learning, 2026. URL https://arxiv.org/abs/2607.01089.

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

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020. URL https://jmlr.org/papers/v21/19-322.html.

Karsten M Borgwardt, Cheng Soon Ong, Stefan Schönauer, SVN Vishwanathan, Alex J Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(suppl_1): i47–i56, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14(64):1884–1899, 2009. doi: 10.1214/EJP.v14-690.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020. URL https://www.jmlr.org/papers/v21/20-163.html.

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

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer, 1996. doi: 10.1007/978-1-4612-0711-5. URL https://luc.devroye.org/pattrec.html.

Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*, 2019. URL https://arxiv.org/abs/1903.02428.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. doi: 10.1016/0097-3165(95)90052-7.

Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1):312–320, 2005. doi: 10.1021/jm040835a.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19620–19631. Curran Associates, Inc., 2020.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34 (5):2326–2366, 2006. doi: 10.1214/009053606000000786. URL https://arxiv.org/abs/math/0702683.

David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003. URL https://www.jmlr.org/papers/v4/mcallester03a.html.

Siqi Miao, Mia Liu, and Pan Li. Interpretable and generalizable graph learning via stochastic attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15524–15543, 2022.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020. URL https://arxiv.org/abs/2007.08663.

Alireza F. Pour and Shai Ben-David. A novel data-dependent learning paradigm for large hypothesis classes. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–27, 2026. URL https://proceedings.mlr.press/v313/pour26a.html.

Kaspar Riesen and Horst Bunke. IAM graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, pp. 287–297. Springer, 2008.

Han Shao, Omar Montasser, and Avrim Blum. A theory of PAC learnability under transformation invariances. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13989–14001. Curran Associates, Inc., 2022. doi: 10.52202/068431-1017.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler–Lehman graph kernels. *Journal of Machine Learning Research*, 12(77): 2539–2561, 2011. URL https://www.jmlr.org/papers/v12/shervashidze11a.html.

Jeffrey J. Sutherland, Lee A. O'Brien, and Donald F. Weaver. Spline-fitting with a genetic algorithm: A method for developing classification structure–activity relationships. *Journal of Chemical Information and Computer Sciences*, 43(6):1906–1915, 2003. doi: 10.1021/ci034143r.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. doi: 10.1214/aos/1079120131.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. Discovering invariant rationales for graph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hGXij5rfiHw.

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

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Xu Zheng, Farhad Shirani, Tianchun Wang, Shouwei Gao, Wenqian Dong, Wei Cheng, and Dongsheng Luo. PAC learnability under explanation-preserving graph perturbations. *arXiv preprint arXiv:2402.05039*, 2024.

A PROOFS OF THE LEARNING GUARANTEES

We write X for the graph input in this appendix. Let $(X, Y) \sim P$, $Y \in \{0, 1\}$, with standard Borel input and key spaces, and let T be a fixed measurable map. Use measurable versions of the conditional probabilities $\eta_x = P(Y = 1 \mid X = x)$ and $\eta_t = P(Y = 1 \mid T = t)$. Write $p_t = P(T = t)$ and

$$\Delta = \mathbb{E}|\eta_X - \eta_{T(X)}|.$$

All class suprema and selected hypotheses below are assumed measurable. The nonempty binary class $\mathcal{H}$ has across-class VC dimension d: a shattered set with pairwise distinct T values has at most d points. Set $\ell(a, b) = b + (1 - 2b)a$ for $a, b \in [0, 1]$.

For $D = ((X_i, Y_i))_{i=1}^m$, $m \geq 1$, define

$$Q_D h(x) = \begin{cases} \dfrac{1}{N_t} \displaystyle\sum_{i:T(X_i)=t} h(X_i), & t = T(x),\ N_t > 0, \\ h(x), & N_t = 0. \end{cases}$$

The operator depends only on D's features. Let $\widehat{h}$ satisfy

$$\widehat{R}_D(Q_D\widehat{h}) \leq \inf_{h \in \mathcal{H}} \widehat{R}_D(Q_D h) + \xi.$$

There is no independent reference set in this definition.

Lemma A.1 (Restricted VC bound)**.** *There is a universal C_0 such that the following holds for every m. Suppose $z_1, \ldots, z_m$ contain at most one distinct point for each T value; repetitions are allowed. For every fixed $b_1, \ldots, b_m \in [0, 1]$ and independent uniform signs σ_i,*

$$\mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \left| \frac{1}{m} \sum_{i=1}^m \sigma_i \ell(h(z_i), b_i) \right| \leq C_0 \sqrt{\frac{d+1}{m}}.$$

Proof. For binary b_i, the binary loss vectors indexed by the sample positions have VC dimension at most d. Two positions with the same representative cannot both be shattered: their loss bits are either equal or complements. Shattering positions from distinct representatives would shatter the corresponding points by $\mathcal{H}$. Apply the usual absolute empirical VC–Rademacher bound (Haussler, 1995; Bartlett & Mendelson, 2002), including the singleton-class term through $d + 1$. For fractional b_i, take independent Bernoulli variables with these means and apply Jensen to the absolute supremum. This proof permits arbitrary repeated representatives and unequal fractional targets. □

Define

$$\begin{aligned} R_0(q) &= \mathbb{E}_X \ell(q(X), \eta_{T(X)}), \\ A(D_X) &= \sup_h \left| R_0(Q_D h) - \frac{1}{m} \sum_i \ell(Q_D h(X_i), \eta_{T(X_i)}) \right|, \\ N(D) &= \sup_h \left| \widehat{R}_D(Q_D h) - \frac{1}{m} \sum_i \ell(Q_D h(X_i), \eta_{X_i}) \right|. \end{aligned}$$

The quantities involving true conditional probabilities are proof devices, not inputs to the algorithm.

Lemma A.2 (Feature deviation with same-sample pooling)**.** *For every $\lambda \geq 0$,*

$$\mathbb{E}e^{\lambda A} \leq \exp\left(2\lambda C_0 \sqrt{\frac{d+1}{m}} + \frac{\lambda^2}{2m} \right).$$

Consequently, with probability at least $1 - \rho$,

$$A \leq 2C_0 \sqrt{\frac{d+1}{m}} + \sqrt{\frac{2\log(1/\rho)}{m}}.$$

Proof. Introduce m independent ghost features X_i', keeping the same operator Q_D in both empirical means. The map $v \mapsto \exp(\lambda\|v\|_\infty)$ is convex. Conditional Jensen therefore gives

$$\mathbb{E}e^{\lambda A} \le \mathbb{E}\exp\left(\lambda \sup_h \left|\frac{1}{m}\sum_i \ell(Q_D h(X_i'), \eta_{T(X_i')}) - \frac{1}{m}\sum_i \ell(Q_D h(X_i), \eta_{T(X_i)})\right|\right).$$

By exchangeability, the two samples on the right can be generated by drawing $2m$ independent features $W_1, \dots, W_{2m}$ and then choosing a uniform m-element subset F of their positions for the fitting role. This auxiliary random split is introduced before conditioning on the ordered pooled features.

Fix the pooled features. For the finitely many keys present, let $I_t = \{j : T(W_j) = t\}$, $k_t = |I_t \cap F|$, and $l_t = |I_t \setminus F|$. Given F, draw representatives independently across keys: r_t is uniform among W_j with $j \in I_t \cap F$ if $k_t > 0$, and uniform among the entire block I_t otherwise. Put $f_h(x) = \ell(h(x), \eta_{T(x)})$. For every h, the empirical difference inside the preceding display is exactly

$$\mathbb{E}_r\left[\frac{1}{m}\sum_t (l_t - k_t) f_h(r_t) \,\middle|\, W, F\right].$$

For a represented key, both empirical means use the fitting average of h. For an unrepresented key, $Q_D h = h$ on its ghost observations, and a uniform ghost representative gives their average. Thus this equality includes unobserved fitting keys. Jensen allows the representative expectation to move outside the exponential supremum.

The crucial decoupling is joint, conditional on the *entire* count vector $(k_t)_t$. Given this vector and W, fitting subsets in different key blocks are independent uniform subsets of the prescribed sizes. For every t, averaging the choice of a uniform representative in that subset makes r_t uniform over I_t, irrespective of k_t (including $k_t = 0$). Hence the representatives are jointly independent uniform block representatives, and their joint law is independent of $(k_t)_t$. The expression above depends on the fitting allocation only through that count vector. Its distribution is therefore unchanged if representatives are drawn first and a fresh uniform balanced allocation is drawn independently.

Generate this fresh allocation by a uniform perfect matching of the $2m$ positions followed by m independent signs, one sign per pair. This gives a uniform balanced allocation by permutation symmetry. Set $z_j = r_{T(W_j)}$. Conditional on W, the representatives, and the pairs (a_i, b_i), its supremum is

$$B = \frac{1}{m}\sup_h \left|\sum_{i=1}^m \sigma_i\{f_h(z_{a_i}) - f_h(z_{b_i})\}\right|.$$

Each of the two lists (z_{a_i}) and (z_{b_i}) has at most one distinct representative per key. Triangle inequality and Lemma A.1 give $\mathbb{E}_\sigma B \le 2C_0\sqrt{(d+1)/m}$. Changing one sign changes B by at most $2/m$, since $f_h \in [0,1]$. The bounded-differences moment-generating-function inequality yields

$$\mathbb{E}_\sigma e^{\lambda B} \le \exp\left(\lambda\mathbb{E}_\sigma B + \frac{\lambda^2}{8}\sum_{i=1}^m (2/m)^2\right).$$

The asserted bound is uniform in every conditioned quantity and follows by integration. Chernoff's bound, optimized at $\lambda = ma$, gives the stated tail with $a = \sqrt{2\log(1/\rho)/m}$. □

Lemma A.3 (Label deviation)**.** *For every $0 < \rho < 1$, with probability at least $1 - \rho$,*

$$N \le 2C_0\sqrt{\frac{d+1}{m}} + \sqrt{\frac{\log(1/\rho)}{2m}}.$$

Proof. Condition on all fitting features. The operator and class of scores are then fixed, and labels are independent Bernoulli variables with possibly unequal means η_{X_i}. Ordinary symmetrization over independent label copies gives

$$\mathbb{E}[N \mid D_X] \le 2\mathbb{E}_{Y,\sigma}\sup_h \left|\frac{1}{m}\sum_i \sigma_i \ell(Q_D h(X_i), Y_i)\right|.$$

Use one uniformly sampled fitting representative per key, shared by all its positions. The pooled loss vector is the expectation of the resulting binary representative loss vector. Jensen and Lemma A.1, conditional on any realized labels and representatives, bound the right-hand side by $2C_0\sqrt{(d+1)/m}$. Changing one fitting label changes N by at most $1/m$. Conditional bounded differences completes the proof; integrating over features preserves the failure probability. No conditional-sufficiency assumption was used. □

Lemma A.4 (Fixed comparator)**.** *For any fixed $h \in \mathcal{H}$ independent of D, with probability at least $1-\rho$,*

$$\widehat{R}_D(Q_D h) \le R_P(h) + \Delta + 3\sqrt{\frac{\log(1/\rho)}{2m}}.$$

Proof. For a key containing $k \ge 1$ examples, put $a = \sum_i Y_i$ and $b = \sum_i h(X_i)$, with the sums restricted to that key. Its contribution to the unnormalized empirical loss is

$$\Phi(k,a,b) = a + b - \frac{2ab}{k} = \frac{1}{k}\sum_{i,j=1}^{k} \ell(h(X_j), Y_i).$$

Let $\mathcal{T}_+ = \{t : p_t > 0\}$ be the countable set of atoms of the key distribution. Independent draws cannot share a non-atomic key: the probability of such a collision is $\int_{\mathcal{T}_+^c} p_t \, dP_T(t) = 0$. Thus non-atomic fitting keys are singletons almost surely, and their pooled and ordinary empirical losses agree.

For $t \in \mathcal{T}_+$, write $\mu_t = \mathbb{E}[h(X) \mid T = t]$ and $C_t = \mathrm{Cov}(h(X), \eta_X \mid T = t)$. Conditional on $N_t = k$, diagonal pairs have expected product $\mathbb{E}[h\eta \mid t]$ and off-diagonal pairs have product $\mu_t\eta_t$. Therefore

$$\mathbb{E}[\Phi \mid N_t = k] = kR_P(h \mid T = t) + 2(k-1)C_t.$$

The contribution is zero for $k = 0$. Since $\mathbb{E}N_t = mp_t$,

$$\begin{aligned}\mathbb{E}\widehat{R}_D(Q_D h) - R_P(h) &= \frac{2}{m}\sum_{t\in\mathcal{T}_+} \mathbb{E}[(N_t-1)_+]C_t \\ &= 2\sum_{t\in\mathcal{T}_+}\left(p_t - \frac{\mathbb{P}(N_t>0)}{m}\right)C_t \le \Delta.\end{aligned}$$

The coefficients in parentheses lie in $[0, p_t]$. Also $2|C_t| \le \mathbb{E}[|\eta_X - \eta_t| \mid T = t]$, since this centered posterior difference has mean zero and $0 \le h \le 1$.

For concentration, adding one example to a group changes Φ by a value in $[-1, 2]$. To see this, let $S \in [0, k^2]$ denote its old pairwise loss sum and let $V \in [0, 2k+1]$ be the new pairs involving the added point. The change is $V/(k+1) - S/[k(k+1)]$, in $[-k/(k+1), (2k+1)/(k+1)]$. The empty-group increment belongs to $[0,1]$. Replacing one sample point is a deletion and an addition, so the normalized total changes by at most $3/m$, even when the old and new keys differ. McDiarmid's inequality now gives the claim. □

Theorem A.5 (Same-sample oracle inequality)**.** *For every $0 < \delta < 1$, with probability at least $1-\delta$,*

$$R_P(Q_D\widehat{h}) \le R^*_{\mathcal{H}} + 3\Delta + \xi + 4C_0\sqrt{\frac{d+1}{m}} + \frac{7}{\sqrt{2}}\sqrt{\frac{\log(4/\delta)}{m}}.$$

Moreover $\Delta \le \sqrt{I(Y;X \mid T)/2}$, with information measured in nats.

Proof. For every realized D and every h,

$$\begin{aligned}R_P(Q_D h) &\le R_0(Q_D h) + \Delta \\ &\le \widehat{R}_D(Q_D h) + A + N + \widehat{\Delta} + \Delta,\end{aligned}$$

where $\widehat{\Delta} = m^{-1}\sum_i |\eta_{X_i} - \eta_{T(X_i)}|$. The inequalities use $|1 - 2Q_D h| \le 1$ and the definitions of A, N. Hoeffding gives $\widehat{\Delta} \le \Delta + \sqrt{\log(1/\rho)/(2m)}$ except with probability ρ. Apply the ERM

inequality to compare $\widehat{h}$ with a fixed h, and apply Lemmas A.2–A.4 with $\rho = \delta/4$. The dimension terms sum to $4C_0\sqrt{(d+1)/m}$, and the four deviation terms sum to $7\sqrt{\log(4/\delta)/(2m)}$. Three approximation contributions each cost at most Δ.

If the comparator infimum is unattained, choose fixed h_j with risks decreasing to it. The exceedance events with thresholds $R_P(h_j)$ plus the common bound increase to the exceedance event at $R^*_{\mathcal{H}}$, so continuity from below preserves probability at most δ without a union bound over j.

Finally, conditional Pinsker gives $|\eta_x - \eta_{T(x)}| \le \sqrt{\mathrm{KL}(\mathrm{Ber}(\eta_x) \| \mathrm{Ber}(\eta_{T(x)}))/2}$. Jensen and the conditional information identity give the last assertion. □

Approximation on recurrent keys. In fact, Theorem A.5 and the risk-dependent bounds in Proposition A.12 and Corollary A.14 hold with Δ replaced by

$$\Delta_+ = \mathbb{E}\big[\mathbf{1}\{T(X) \in \mathcal{T}_+\}|\eta_X - \eta_{T(X)}|\big].$$

To see this, set $\widetilde{\eta}_x = \eta_{T(x)}$ on atomic keys and $\widetilde{\eta}_x = \eta_x$ elsewhere. Use $\mathbb{E}\ell(q(X), \widetilde{\eta}_X)$ as the auxiliary risk. Non-atomic keys are singletons in the pooled fitting and ghost samples, and a fresh query almost surely never shares such a key with fitting data. Therefore the identity $Q_D\ell(h, \widetilde{\eta}) = \ell(Q_D h, \widetilde{\eta})$ holds on all observations used by the feature arguments, almost surely. The representative proofs remain unchanged. The empirical posterior discrepancy is now the average of $\mathbf{1}\{T(X_i) \in \mathcal{T}_+\}|\eta_{X_i} - \eta_{T(X_i)}|$, and the comparator covariance sum already involves only atoms. Thus each approximation step costs Δ_+ instead of Δ.

Writing $\kappa_+ = \sum_{t \in \mathcal{T}_+} p_t I(Y; X \mid T = t)$, conditional Pinsker gives $\Delta_+ \le \sqrt{P(T \in \mathcal{T}_+)\kappa_+/2}$. For a non-atomic key distribution this penalty is zero: pooling leaves fitting and fresh-query predictions unchanged almost surely, and distinct-key sample complexity suffices. The finite-key experiments have $\Delta_+ = \Delta$.

Weighted majority-label formulation. For $\bar{Y}_t = N_t^{-1} \sum_{i:T(X_i)=t} Y_i$, linearity gives

$$\widehat{R}_D(Q_D h) = \frac{1}{m} \sum_i \ell(h(X_i), \bar{Y}_{T(X_i)}).$$

Let $c_t = \mathbf{1}\{\bar{Y}_t > 1/2\}$, $w_t = |1 - 2\bar{Y}_t|$. Then the objective is

$$\frac{1}{m} \sum_i \min\{\bar{Y}_{T(X_i)}, 1 - \bar{Y}_{T(X_i)}\} + \frac{1}{m} \sum_i w_{T(X_i)} \mathbf{1}\{h(X_i) \ne c_{T(X_i)}\}.$$

Thus the proposed learner is exact ordinary weighted ERM on group-majority pseudo-labels. Groups tied at one half have weight zero. This identity makes no independence assumption and does not create additional independent labels.

Deterministic output. For the score $q = Q_D\widehat{h}$, define $g_u = \mathbf{1}\{q \ge u\}$, $0 \le u \le 2$. Layer cake gives $\int_0^1 R_P(g_u)\, du = R_P(q)$, hence $\inf_u R_P(g_u) \le R_P(q)$. Conditional on D, a fresh validation sample of size v obeys

$$\mathbb{P}\{\sup_u |\widehat{R}_V(g_u) - R_P(g_u)| > 3a \mid D\} \le 6e^{-2va^2}.$$

To see this, for $y \in \{0, 1\}$ define $Z_y = q(X)$ when $Y = y$ and $Z_y = 3$ otherwise, and write $F_y^<(u) = \mathbb{P}(Z_y < u)$. For $u \in [0, 2]$,

$$R_P(g_u) = \mathbb{P}(Y = 0) + F_1^<(u) - F_0^<(u).$$

The same identity holds empirically. Apply DKW (Massart, 1990) to each Z_y and Hoeffding to the class prior. A union bound gives the displayed tail; strict left limits handle threshold ties. An empirical minimizer over validation thresholds exists because there are finitely many distinct loss vectors. Its population error is at most $\inf_u R_P(g_u) + 6a$. Use $a = \epsilon/12$ and $v = \lceil 72 \log(12/\delta)/\epsilon^2 \rceil$, so validation costs at most $\epsilon/2$ with failure probability $\delta/2$. Choose m so that Theorem A.5, with confidence $\delta/2$, costs at most $\epsilon/2$ (including the optimization tolerance, or take exact ERM). The final deterministic predictor then satisfies

$$R_P(\widehat{g}) \le R^*_{\mathcal{H}} + 3\sqrt{I(Y; X \mid T)/2} + \epsilon$$

with probability at least $1-\delta$, using $m+v = O((d+\log(1/\delta))/\epsilon^2)$ labeled observations for $\delta < 1/2$. Under exact sufficiency, the approximation term is zero even with arbitrary intrinsic label noise. This result concerns statistical learnability; arbitrary weighted ERM need not be computationally efficient.

Optimality at $\kappa = 0$. The dependence on d/ϵ^2 cannot be improved uniformly over distributions with $I(Y; X \mid T) = 0$. If T is injective on the support, then $I(Y; X \mid T) = 0$ for every conditional label distribution and the across-class dimension equals $\mathrm{VC}(\mathcal{H})$. The problem is then ordinary agnostic classification with the same comparator, and the minimax lower bounds for arbitrary classification rules (Devroye et al., 1996, Chapter 14) require order $\mathrm{VC}(\mathcal{H})/\epsilon^2$ observations. The same order remains necessary when labels are deterministic functions of the input, by an agnostic lower bound (Shao et al., 2022, Theorem 8). Theorem A.5 and the deterministic-output bound therefore match these lower bounds up to constants at $\kappa = 0$, while allowing arbitrary label noise within each key.

A.1 Fast Rates and Decision Sufficiency

Define $\Pi_d(n) = \sum_{j=0}^{\min\{d,n\}} \binom{n}{j}$, with $\Pi_0(n) = 1$. **Feature offset.**

Lemma A.6. *Let $f_0, f_1 : \mathcal{X} \to [0,1]$ be any fixed measurable functions. Put $f_h(x) = f_{h(x)}(x)$ and define*

$$M(h) = \mathbb{E}_X Q_D f_h(X), \qquad \bar{M}(h) = \frac{1}{m}\sum_i Q_D f_h(X_i).$$

Here Q_D acts linearly on real-valued functions using the same definition. For every $\alpha > 0$, with $\lambda = 4\alpha m/(2+\alpha)^2$,

$$\mathbb{E}\exp\big(\lambda \sup_h\{M(h) - (1+\alpha)\bar{M}(h)\}\big) \le \Pi_d(2m).$$

Consequently, with probability at least $1-\rho$, simultaneously for all h,

$$M(h) \le (1+\alpha)\bar{M}(h) + \frac{(2+\alpha)^2}{4\alpha m}\log\frac{\Pi_d(2m)}{\rho}.$$

Proof. Conditional Jensen with an independent ghost feature sample $(X_i')_{i=1}^m$ gives an upper bound by the expected exponential supremum of

$$\frac{1}{m}\sum_i Q_D f_h(X_i') - \frac{1+\alpha}{m}\sum_i Q_D f_h(X_i).$$

By exchangeability, draw $2m$ independent features W and a uniform m-element fitting subset F before conditioning on W. For each key present let k_t and l_t be its fitting and ghost counts. Given W, F, draw a common representative r_t uniformly from its fitting positions if $k_t > 0$, and uniformly from its ghost positions otherwise. The displayed expression is exactly the representative expectation of

$$\frac{1}{m}\sum_t \{l_t - (1+\alpha)k_t\} f_h(r_t).$$

For $k_t > 0$ this follows from the defining fitting average; for $k_t = 0$, the operator leaves ghost predictions unpooled and their sum equals l_t times their average. Convexity of the exponential supremum moves the representative expectation outside.

Given the entire vector $(k_t)_t$ and the pooled features, the fitting subsets in different key blocks are independent uniform subsets of their prescribed sizes. Averaging a uniform fitting representative makes each r_t uniform over the entire pooled block, independently across blocks. The same is true when $k_t = 0$. Thus the joint representative law is independent of the entire count vector. The last expression depends on F only through those counts, so it has the same distribution if representatives are selected first, independently of a fresh uniform balanced allocation.

Generate that allocation by a uniform perfect matching of the pooled positions and independent signs σ_i. Conditional on the pooled features, representatives, and matching, let $a_i = f_h(r_{T(W_{u_i})})$ and $b_i = f_h(r_{T(W_{v_i})})$ for pair (u_i, v_i). Its unnormalized contribution is

$$\frac{2+\alpha}{2}\sigma_i(a_i - b_i) - \frac{\alpha}{2}(a_i + b_i).$$

For fixed h, independence of signs, $\cosh z \le e^{z^2/2}$, and $(a_i - b_i)^2 \le a_i + b_i$ for $a_i, b_i \in [0,1]$ imply that the moment generating function is at most

$$\exp\left[\left\{\frac{\lambda^2(2+\alpha)^2}{8m^2} - \frac{\lambda\alpha}{2m}\right\}\sum_i (a_i + b_i)\right] = 1.$$

On the set of at most $2m$ distinct representatives, all keys are different. Sauer's lemma gives at most $\Pi_d(2m)$ binary prediction patterns. Applying the fixed coordinatewise maps f_0, f_1 cannot increase this number. Therefore the exponential of the supremum is bounded by the sum of at most $\Pi_d(2m)$ exponentials, each with conditional expectation at most one. Integrating proves the moment bound. Markov's inequality gives the tail. □

Posterior sufficiency and Massart noise. Assume $Y \perp X \mid T$, put $\eta_t = \mathbb{P}(Y = 1 \mid T = t)$, and suppose $\gamma_t = |2\eta_t - 1| \ge \gamma > 0$ for P_T-almost every key. Suppose a Bayes rule $h^* \in \mathcal{H}$. It agrees almost surely with $\mathbf{1}\{\eta_{T(X)} > 1/2\}$. A single probability-one event for the finite samples and test distribution suffices for all identities below. Let $R^* = R(h^*)$ and

$$e_h = \mathbf{1}\{h \ne h^*\}, \qquad f_h = \gamma_T e_h, \qquad \mathcal{E}(h) = R(Q_D h) - R^*,$$
$$\bar{\mathcal{E}}(h) = \frac{1}{m}\sum_i \gamma_{T(X_i)} |Q_D h(X_i) - h^*(X_i)|.$$

$$\widehat{\mathcal{E}}(h) = \widehat{R}_D(Q_D h) - \widehat{R}_D(h^*).$$

Keywise constancy of h^* implies $|Q_D h - h^*| = Q_D e_h$ and $\mathcal{E}(h) = \mathbb{E} Q_D f_h(X)$, including unseen keys. Thus Lemma A.6 applies directly.

Lemma A.7. *For $\beta > 0$ and $\lambda = 2\beta\gamma m/(1+\beta)^2$,*

$$\mathbb{E}\left[\exp\left(\lambda \sup_h \{\bar{\mathcal{E}}(h) - (1+\beta)\widehat{\mathcal{E}}(h)\}\right) \,\middle|\, D_X\right] \le \Pi_d(m).$$

Hence, with probability at least $1 - \rho$, simultaneously for all h,

$$\bar{\mathcal{E}}(h) \le (1+\beta)\widehat{\mathcal{E}}(h) + \frac{(1+\beta)^2}{2\beta\gamma m}\log\frac{\Pi_d(m)}{\rho}.$$

Proof. Given D_X, the variables $Z_i = (1 - 2Y_i)(1 - 2h^*(X_i))$ are independent, belong to $\{-1, 1\}$, and have means $\gamma_{T(X_i)}$. The empirical excess equals $m^{-1}\sum_i Z_i Q_D e_h(X_i)$. Select a common uniform fitting representative per observed key, independently of all labels. Jensen bounds the exponential supremum by its expectation after replacing every $Q_D e_h(X_i)$ by $e_h(r_{T(X_i)})$. Given the representatives, at most $\Pi_d(m)$ distinct binary disagreement vectors occur. For any fixed such vector e, conditional Hoeffding bounds its exponential moment by

$$\exp\left(\sum_i e_i\left\{-\frac{\lambda\beta\gamma_{T(X_i)}}{m} + \frac{\lambda^2(1+\beta)^2}{2m^2}\right\}\right) \le 1.$$

The inequality uses $\gamma_t \ge \gamma$ and the chosen value of λ; the squared coefficient is $e_i^2 = e_i$. Summing over patterns and integrating over the representatives proves the conditional moment bound. Markov and integration over D_X give the tail. □

Theorem A.8. *Under these assumptions, with probability at least $1 - \delta$,*

$$R(\hat{q}) - R^* \le 4\xi + \frac{25}{4\gamma m}\log\frac{2\Pi_d(2m)}{\delta}.$$

On the same event, $\hat{g} = \mathbf{1}\{\hat{q} \ge 1/2\}$ satisfies

$$R(\hat{g}) - R^* \le 8\xi + \frac{25}{2\gamma m}\log\frac{2\Pi_d(2m)}{\delta}.$$

Neither fitting nor thresholding requires knowledge of γ or fresh validation labels.

Proof. Since $Q_D h^* = h^*$ almost surely and $h^* \in \mathcal{H}$, optimization gives $\widehat{\mathcal{E}}(\hat{h}) \le \xi$. Use $\alpha = \beta = 1$ in the two lemmas, each with failure probability $\delta/2$. Their intersection gives

$$\mathcal{E}(\hat{h}) \le 4\xi + \frac{4}{\gamma m} \log \frac{2\Pi_d(m)}{\delta} + \frac{9}{4m} \log \frac{2\Pi_d(2m)}{\delta}.$$

Use $\gamma \le 1$ and monotonicity of Π_d for the claimed constant. Finally $\mathbf{1}\{\hat{g} \ne h^*\} \le 2|\hat{q} - h^*|$ pointwise, and multiply by the conditional margin and integrate. □

For $1 \le d \le 2m$, $\log \Pi_d(2m) \le d \log(2em/d)$. For $d = 0$ use $\Pi_0 = 1$; for $d > 2m$ the unsimplified bound remains valid. If $\gamma = 1$, then $R^* = 0$ and $\widehat{\mathcal{E}} = \bar{\mathcal{E}}$ exactly, so the feature lemma alone gives $R(\hat{q}) \le 2\xi + 9 \log(\Pi_d(2m)/\delta)/(4m)$, and the hard-threshold risk is at most twice this value.

Adaptation to a margin distribution. A uniform positive margin can be replaced by control of nearly ambiguous keys. Assume $Y \perp X \mid T$ and a keywise constant Bayes rule $h^* \in \mathcal{H}$. No positive lower bound on $\gamma_T = |2\eta_T - 1|$ is imposed. For $0 < \theta \le 1$, define

$$B(\theta) = \mathbb{E}[(\theta - \gamma_T)_+], \qquad \widehat{B}(\theta) = \frac{1}{m} \sum_i (\theta - \gamma_{T(X_i)})_+, \quad a = \frac{1}{m} \log \frac{3\Pi_d(2m)}{\delta}.$$

Proposition A.9 (Margin-distribution bound)**.** *For the same pooled ERM, with probability at least* $1 - \delta$*,*

$$R(\hat{q}) - R^* \le 4\xi + \inf_{0 < \theta \le 1} \left\{ \left(\frac{4}{\theta} + 2\theta + \frac{9}{4} \right) a + 4B(\theta) \right\}.$$

Half-thresholding doubles this bound. Neither output requires knowledge of the margin distribution or fresh validation labels.

Proof. Fix θ and condition on all fitting features. In the label-offset proof, use $\beta = 1$ and $\lambda = \theta m/2$. For a fixed representative disagreement pattern e, its log moment is at most

$$\sum_i e_i \left\{ -\frac{\lambda \gamma_{T(X_i)}}{m} + \frac{2\lambda^2}{m^2} \right\} = \frac{\lambda}{m} \sum_i e_i (\theta - \gamma_{T(X_i)}) \le \lambda \widehat{B}(\theta).$$

The same representative averaging and pattern count therefore give

$$\mathbb{E}\left[e^{\lambda \sup_h \{\bar{\mathcal{E}}(h) - 2\widehat{\mathcal{E}}(h) - \widehat{B}(\theta)\}} \,\middle|\, D_X \right] \le \Pi_d(m).$$

Thus $\bar{\mathcal{E}}(h) \le 2\widehat{\mathcal{E}}(h) + \widehat{B}(\theta) + 2\log(\Pi_d(m)/\rho)/(\theta m)$ uniformly in h, except with probability ρ. The variables $(\theta - \gamma_{T(X_i)})_+/\theta$ are iid in $[0, 1]$. Convexity and $e - 1 < 2$ give their exponential-moment bound and hence

$$\widehat{B}(\theta) \le 2B(\theta) + \frac{\theta}{m} \log \frac{1}{\rho}$$

except with probability ρ. Use the feature-offset lemma with parameter one, allocate $\rho = \delta/3$ to these three events, and apply $\widehat{\mathcal{E}}(\hat{h}) \le \xi$. This yields

$$R(\hat{q}) - R^* \le 4\xi + 4B(\theta) + \left(\frac{4}{\theta} + 2\theta + \frac{9}{4} \right) a.$$

The learner is independent of the proof parameter θ. A deterministic sequence of θ values approaching the infimum of the displayed bound, and continuity of increasing exceedance events, gives the infimum without a union over θ. The same pointwise half-threshold inequality as in Theorem A.8 completes the proof. □

For example, suppose for some $A \ge 1$ and $\alpha > 0$ that

$$\mathbb{P}\{\gamma_T \le u\} \le A u^\alpha \qquad (0 < u \le 1). \tag{19}$$

This is a classical polynomial low-noise condition (Tsybakov, 2004). It excludes a positive mass of ties, so exact sufficiency makes every Bayes rule keywise constant almost surely. Since $B(\theta) = \int_0^\theta \mathbb{P}(\gamma_T < u)\, du \le A\theta^{\alpha+1}/(\alpha + 1)$, Proposition A.9 implies

$$R(\hat{g}) - R^* \le 8\xi + 25 A^{1/(\alpha+2)} \left(\frac{\log(3\Pi_d(2m)/\delta)}{m} \right)^{(\alpha+1)/(\alpha+2)}. \tag{20}$$

Indeed, when $a \le A$, choose $\theta = (a/A)^{1/(\alpha+2)}$. The parenthetical coefficient is at most $33/(4\theta)$, giving $49/4$ for the score and $49/2 < 25$ after thresholding. When $a > A \ge 1$, the bound follows from excess risk at most one. The margin-tail rate is classical; the result here is that the same empirically pooled learner attains it with across-class complexity and without margin tuning.

Decision sufficiency with heterogeneous noise. A weaker assumption is that a Bayes rule $h^* \in \mathcal{H}$ satisfies $h^* = c \circ T$ almost surely for a measurable binary c, while $\gamma_x = |2\mathbb{P}(Y = 1 \mid X = x) - 1| \ge \gamma > 0$ almost surely. The posterior may vary within a key. Define

$$M(h) = \mathbb{E}Q_D e_h(X), \qquad \bar{M}(h) = \frac{1}{m}\sum_i Q_D e_h(X_i).$$

The feature lemma with $f_h = e_h$ applies. The label excess still equals $m^{-1}\sum_i Z_i Q_D e_h(X_i)$, but now $\mathbb{E}[Z_i \mid D_X] = \gamma_{X_i}$. For $\beta > 0$ and $\lambda = 2\beta\gamma^2 m/(1+\beta)^2$, the proof of Lemma A.7 gives

$$\mathbb{E}\left[\exp\left(\lambda \sup_h \left\{\bar{M}(h) - \frac{1+\beta}{\gamma}\widehat{\mathcal{E}}(h)\right\}\right) \,\middle|\, D_X\right] \le \Pi_d(m).$$

Indeed each fixed binary pattern has log moment at most

$$\sum_i e_i \left\{\frac{\lambda}{m}\left(1 - \frac{(1+\beta)\gamma_{X_i}}{\gamma}\right) + \frac{\lambda^2(1+\beta)^2}{2\gamma^2 m^2}\right\} \le 0.$$

Combining the resulting tail with the feature lemma, $\alpha = \beta = 1$, and ERM gives, with probability at least $1 - \delta$,

$$M(\hat{h}) \le \frac{4\xi}{\gamma} + \frac{25}{4\gamma^2 m}\log\frac{2\Pi_d(2m)}{\delta}.$$

Because $R(\hat{q}) - R^* = \mathbb{E}[\gamma_X |\hat{q}(X) - h^*(X)|] \le M(\hat{h})$, this proves a fast risk bound under decision sufficiency alone. Half-thresholding doubles the bound. The price here is weaker γ^{-2} dependence; no claim of optimality is made. This argument does not require small $I(Y; X \mid T)$.

A.2 Risk-Dependent Rates without Bayes Realizability

We allow arbitrary P and retain the same pooled ERM and fresh-threshold validation. No positive margin or Bayes classifier in $\mathcal{H}$ is assumed. Write

$$F_h = \widehat{R}_D(Q_D h), \qquad L_h = \mathbb{E}[F_h \mid D_X] = \frac{1}{m}\sum_i \ell(Q_D h(X_i), \eta_{X_i}),$$

$$L_h^0 = \frac{1}{m}\sum_i \ell(Q_D h(X_i), \eta_{T(X_i)}).$$

The feature-offset lemma applies to $f_h(x) = \ell(h(x), \eta_{T(x)})$, giving an upper bound on $R_0(Q_D h)$ in terms of L_h^0. Below, the label offset uses the actual posteriors η_{X_i}.

Lemma A.10 (Nonnegative label-loss offset)**.** *For $\beta > 0$ and $\lambda = 2\beta m/(1+\beta)^2$,*

$$\mathbb{E}\left[\exp\left(\lambda \sup_h\{L_h - (1+\beta)F_h\}\right) \,\middle|\, D_X\right] \le \Pi_d(m).$$

Thus, with failure probability at most ρ, simultaneously for all h,

$$L_h \le (1+\beta)F_h + \frac{(1+\beta)^2}{2\beta m}\log\frac{\Pi_d(m)}{\rho}.$$

Proof. Given D_X, draw a common uniform fitting representative per key independently of the labels. Both L_h and F_h are expectations over this choice. Jensen moves it outside the exponential supremum. For a fixed representative prediction pattern, the resulting binary losses $B_i = \ell(h(r_{T(X_i)}), Y_i)$ are independent with means $\mu_i = \ell(h(r_{T(X_i)}), \eta_{X_i})$. Put $z = \lambda(1+\beta)/m$. Then

$$\begin{aligned}\log\mathbb{E}\exp\left(\frac{\lambda}{m}\sum_i\{\mu_i - (1+\beta)B_i\}\right) &\le \sum_i \mu_i\left\{\frac{\lambda}{m} + e^{-z} - 1\right\} \\ &\le \sum_i \mu_i\left\{-\frac{\lambda\beta}{m} + \frac{\lambda^2(1+\beta)^2}{2m^2}\right\} = 0.\end{aligned}$$

Here $\log(1+u) \le u$ and $e^{-z} \le 1 - z + z^2/2$ for $z \ge 0$. At most $\Pi_d(m)$ representative patterns occur. Sum their moments, integrate over the representatives, and apply Markov. □

Lemma A.11 (Relative concentration of a pooled comparator)**.** *Fix h independently of D. For every $0 < u \le 1$, with probability at least $1 - \rho$,*

$$F_h \le (1+u)\{R(h) + \Delta\} + \frac{2(1+u)}{um} \log \frac{1}{\rho}.$$

Proof. Let $Z = mF_h$, and let Z_i be the unnormalized pooled loss after deleting observation i, keeping h fixed. Thus Z_i is measurable in all observations except i. We show the sharp squared-deletion inequality

$$-1 \le Z - Z_i \le 2, \qquad \sum_{i=1}^{m} (Z - Z_i)^2 \le 2Z. \tag{21}$$

The range follows from Lemma A.4. For the squared bound, consider one group of size $k \ge 2$, with positive-label count a and positive-prediction count b. Its loss is $\Phi = a + b - 2ab/k$. Let $c_{00}, c_{01}, c_{10}, c_{11}$ count the four pairs $(h(X_i), Y_i)$, and let $D_{00}, D_{01}, D_{10}, D_{11}$ be their deletion changes. Substituting in Φ gives

$$\begin{aligned} k(k-1)D_{00} &= 2ab, & k(k-1)D_{11} &= 2(k-a)(k-b), \\ k(k-1)D_{01} &= k(k-1) - 2b(k-a), & k(k-1)D_{10} &= k(k-1) - 2a(k-b). \end{aligned}$$

Put $v = c_{01} + c_{10}$, $w = c_{01}c_{10}$, and $z = c_{00}c_{11} - w$. Expanding the four squares yields

$$\begin{aligned} k(k-1)^2 \left(2\Phi - \sum_{\alpha,\beta \in \{0,1\}} c_{\alpha\beta} D_{\alpha\beta}^2 \right) &= kv(k-1)^2 + 4wk(k-2) \\ &\quad + 4z(v-1)(2k-1) + 12z^2 =: P_k. \end{aligned}$$

If $v = 0$, then $w = 0$ and $z = c_{00}c_{11}$. Either $z = 0$, or the two positive integer counts sum to k and hence $z \ge k - 1$. Thus $P_k = 4z\{3z - (2k-1)\} \ge 0$. If $v \ge 1$ and $z \ge 0$, every displayed term of P_k is nonnegative.

It remains to consider $v \ge 1$ and $z < 0$. Set $e = -z \le w$ and $B = \{(2k-1)v - k^2 + 1\}_+$. Completing the square gives

$$\begin{aligned} P_k &\ge kv(k-1)^2 + 4e\{k^2 - 1 - (2k-1)v\} + 12e^2 \\ &\ge kv(k-1)^2 - B^2/3. \end{aligned}$$

Since $v \le k$ and $k \ge 2$, $B \le v(k - 1 + 1/k) \le (3/2)v(k-1)$. Therefore $B^2 \le (9/4)kv(k-1)^2$, proving $P_k \ge 0$. A singleton has deletion change $\Phi \in \{0, 1\}$. Summing the group bounds proves (21). Its factor two is attained by a two-point group with pairs $(0, 0)$ and $(1, 1)$.

For completeness, apply the entropy method with these deletion statistics (Boucheron et al., 2009, Theorem 2). If $\phi(v) = e^v - v - 1$, entropy tensorization and its variational formula give

$$\mathrm{Ent}(e^{\lambda Z}) \le \mathbb{E}\left[e^{\lambda Z} \sum_i \phi(-\lambda(Z - Z_i)) \right].$$

This inequality does not require $Z_i \le Z$. For $0 < \lambda < 1/2$, the arguments have magnitude below one, and $\phi(v) \le v^2$ on $[-1, 1]$. Hence the right-hand side is at most $2\lambda^2 \mathbb{E}[Ze^{\lambda Z}]$. Writing $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z}$ yields

$$\lambda(1-2\lambda)\psi'(\lambda) \le \psi(\lambda), \qquad \left(\frac{1-2\lambda}{\lambda} \psi(\lambda) \right)' \le 0.$$

The expression inside the derivative tends to $\mathbb{E}Z$ as $\lambda \downarrow 0$, so $\psi(\lambda) \le \lambda \mathbb{E}Z/(1-2\lambda)$. Set $\lambda = u/[2(1+u)]$ and apply Markov. Lemma A.4's expectation identity gives $\mathbb{E}F_h \le R(h) + \Delta$, completing the proof. In particular, this step makes no posterior-sufficiency assumption. □

Proposition A.12 (Risk-dependent pooled oracle)**.** *Let $a = m^{-1}\log(4\Pi_d(2m)/\delta)$. For arbitrary P, with probability at least $1-\delta$,*

$$R(\widehat{q}) \le R^*_{\mathcal{H}} + 3\Delta + 4\xi + 35\sqrt{(R^*_{\mathcal{H}} + \Delta)a} + 35a.$$

Proof. Fix h and $u \in (0,1]$. Use $\alpha = \beta = u$ in Lemmas A.6 and A.10, and apply Lemma A.11 to this comparator. A fourth event controls $\widehat{\Delta} = m^{-1}\sum_i |\eta_{X_i} - \eta_{T(X_i)}|$:

$$\widehat{\Delta} \le (1+u)\Delta + \frac{\log(1/\rho)}{um}.$$

Indeed, its summands W_i lie in $[0,1]$ with mean Δ; convexity gives $\mathbb{E}e^{uW_i} \le 1 + \Delta(e^u - 1) \le e^{(u+u^2)\Delta}$ for $0 < u \le 1$, followed by Markov.

Allocate failure probability $\delta/4$ to each event. For every h', $R(Q_D h') \le R_0(Q_D h') + \Delta$ and $L^0_{h'} \le L_{h'} + \widehat{\Delta}$. Combining the four events with empirical minimization gives

$$\begin{aligned} R(\widehat{q}) &\le \Delta + (1+u)^3\{r_h + \Delta\} + (1+u)^2\Delta + (1+u)^2\xi \\ &\quad + \left\{2(1+u)^3 + \frac{(1+u)^3}{2} + \frac{(2+u)^2}{4} + (1+u)\right\}\frac{a}{u} \\ &\le r_h + 3\Delta + 4\xi + 10u(r_h + \Delta) + \frac{97a}{4u}. \end{aligned}$$

Choose $u = \min\{1, \sqrt{a/(r_h+\Delta)}\}$, with $u = 1$ at a zero denominator. If $r_h + \Delta \ge a$, the last two terms are at most $(137/4)\sqrt{(r_h+\Delta)a}$; otherwise they are at most $(137/4)a$. This proves the looser displayed bound. A decreasing sequence of fixed comparator risks and continuity of their exceedance events handles an unattained infimum, without a union over comparators. The optimized proof parameter is not an input to the learner. □

Lemma A.13 (Relative-risk threshold validation)**.** *Fix a measurable score $q : \mathcal{X} \to [0,1]$ independently of an iid validation sample of size v. Let $\widehat{g}$ minimize empirical loss over $g_u = \mathbf{1}\{q \ge u\}$, $u \in [0,2]$. Put $s = \inf_u R(g_u)$ and $b = v^{-1}\log(2(2v+1)/\delta)$. With probability at least $1-\delta$,*

$$R(\widehat{g}) \le s + 8\sqrt{sb} + 8b.$$

Proof. On n labeled observations the threshold loss vectors have at most $n+1$ patterns: XOR with the labels preserves the count of the nested prediction vectors. The ordinary version of the feature-offset argument gives, uniformly in u and with failure probability ρ,

$$R(g_u) \le (1+\alpha)\widehat{R}_V(g_u) + \frac{(2+\alpha)^2}{4\alpha v}\log\frac{2v+1}{\rho}.$$

To obtain it, introduce a ghost validation sample, condition on the pooled $2v$ labeled observations, and generate their roles with a random pairing and independent signs. The pair contribution and exponential-moment calculation are identical to Lemma A.6, now applied directly to binary loss values without a pooling operator. Sum over at most $2v+1$ patterns. This argument does not require a countable score range.

For a fixed population comparator threshold, the Bernoulli moment calculation gives $\widehat{R}_V(g) \le (1+z)R(g) + \log(1/\rho)/(zv)$ for $0 < z \le 1$. Set $\alpha = z = u \in (0,1]$, use $\rho = \delta/2$, and apply empirical minimization:

$$R(\widehat{g}) \le R(g) + 3uR(g) + \frac{17b}{4u}.$$

Optimizing $u = \min\{1, \sqrt{b/R(g)}\}$ and taking a decreasing sequence of comparator risks gives the claim, with no union over thresholds. A finite validation sample induces finitely many loss patterns, so its empirical minimum exists. □

Corollary A.14 (Deterministic risk-dependent oracle)**.** *Use m fitting labels and m fresh validation labels for the same threshold ERM. Set*

$$z = \frac{1}{m} \max \left\{ \log \frac{8\Pi_d(2m)}{\delta},\ \log \frac{4(2m+1)}{\delta} \right\}.$$

For arbitrary P, with probability at least $1-\delta$,

$$R(\widehat{g}) \le R^*_{\mathcal{H}} + 3\Delta + 12\xi + 51\sqrt{(R^*_{\mathcal{H}} + \Delta)z} + 123z.$$

For exact ERM, the tolerance term disappears and $123z$ can be replaced by $115z$. The total label count is $2m$.

Proof. Apply Proposition A.12 and, conditional on the fitting sample, Lemma A.13, each with failure probability $\delta/2$. Layer cake gives $s \le R(\widehat{q})$. Put $t = R^*_{\mathcal{H}} + \Delta$. The fitting bound implies

$$\sqrt{R(\widehat{q})} \le \sqrt{3t + 35\sqrt{tz} + 35z + 4\xi} \le 2\sqrt{t} + 9\sqrt{z} + 2\sqrt{\xi}.$$

Substitute this into the validation bound and use $16\sqrt{\xi z} \le 8\xi + 8z$. The square-root coefficient is $35 + 16 = 51$ and the linear coefficient is $35 + 72 + 8 + 8 = 123$. At $\xi = 0$ the last $8z$ is unnecessary. Finally, $\log \Pi_d(2m) \le d \log(2em)$ gives the main text's universal-constant form. The learner needs neither $R^*_{\mathcal{H}}$ nor Δ. □

The estimation term improves when both the comparator risk and the posterior discrepancy are small. The margin theorem instead controls excess above a possibly large Bayes risk. This argument pays growth-function logarithms; Theorem 3.1 retains its log-free square-root rate.

Conditional averaging under exact sufficiency. Exact posterior sufficiency gives a further interpretation of the fixed-comparator loss and a sharper concentration argument.

Lemma A.15 (Conditional averaging of a fixed comparator)**.** *Assume $Y \perp X \mid T$. For any fixed h independent of D, write $r_h = R(h)$. For every real z,*

$$\mathbb{E}e^{zmF_h} \le (1 - r_h + r_h e^z)^m.$$

For $0 < z \le 1$, with probability at least $1-\rho$,

$$F_h \le (1+z)r_h + \frac{\log(1/\rho)}{zm}.$$

Proof. Condition on all fitting features and the positive-label count in every key. Posterior sufficiency makes all assignments of those labels to positions within their key equally likely, independently across keys. Averaging the ordinary empirical loss over these assignments yields F_h exactly. Conditional Jensen bounds its exponential moment by that of ordinary empirical loss, an average of iid Bernoulli variables with mean r_h. For $0 < z \le 1$, use $e^z \le 1 + z + z^2$ and $\log(1+u) \le u$ to obtain $\mathbb{E}\exp\{zm(F_h - (1+z)r_h)\} \le 1$, then apply Markov. This conditional-averaging argument concerns a fixed comparator; fitted hypotheses require the uniform bounds above. □

The averaging also gives an exact variance identity. Put $\mu_t = \mathbb{P}(h(X) = 1 \mid T = t)$. Then

$$\mathrm{Var}(F_h) = \frac{r_h(1-r_h)}{m} - \frac{4}{m^2} \sum_{t:p_t>0} \mathbb{E}[(N_t - 1)_+]\eta_t(1-\eta_t)\mu_t(1-\mu_t). \tag{22}$$

Indeed, conditional on features and a key's label count, write $k = N_t$, $a = \sum Y_i$, $b = \sum h(X_i)$. For $k \ge 2$, $C = \sum h(X_i)Y_i$ has hypergeometric variance $a(k-a)b(k-b)/(k^2(k-1))$, while the unpooled error count is $a + b - 2C$. Average this conditional variance over the independent label and prediction counts within each key, then use total variance. Keys with fewer than two observations contribute zero, including every non-atomic key almost surely. Thus repeated keys reduce the variance of F_h relative to ordinary empirical loss when both predictions and labels vary within a key. Conditional averaging and its variance benefit are classical (Chen et al., 2020); equation (22) describes this empirical label-permutation operation.

A.3 Selective Majority Sharing

This argument permits arbitrary posterior distributions. Here T is fixed before sampling and takes finitely many values almost surely. Let $c_t^* = \mathbf{1}\{\eta_t > 1/2\}$ (either label is valid at a tie), $\gamma_t = |2\eta_t - 1|$, $r^*(x) = \min\{\eta_x, 1-\eta_x\}$, and $r_t^* = \min\{\eta_t, 1-\eta_t\}$. For a realized baseline b and set of selected keys A, define

$$V_A(b) = \mathbb{E}_X\big[\mathbf{1}\{T(X) \in A\}\{\ell(b(X), \eta_X) - r^*(X)\}\big],$$
$$\Gamma_T(A) = \sum_{t \in A} p_t\{r_t^* - \mathbb{E}[r^*(X) \mid T = t]\}.$$

These expectations concern a fresh query, conditional on the complete fitted output. Both quantities are nonnegative: the first by Bayes optimality and the second by concavity of $u \mapsto \min\{u, 1-u\}$. The selected set and baseline may depend on fitting and additional development data; both are fixed before the fresh query is drawn.

Let $\hat{c}_t$ be the empirical majority in every observed key, resolving ties arbitrarily. The set A selects only observed keys with a unique majority, and h_A overrides b on A. On each selected class, the excess of $\hat{c}_t$ over its Bayes-T label is exactly $\gamma_t \mathbf{1}\{\hat{c}_t \neq c_t^*\}$. Inserting the Bayes-$T$ and unrestricted Bayes risks gives the exact identity

$$R(h_A) - R(b) = -V_A(b) + \Gamma_T(A) + \sum_{t \in A} p_t \gamma_t \mathbf{1}\{\hat{c}_t \neq c_t^*\}. \tag{23}$$

Outside A the two predictors agree. This is a pointwise identity in the training output, so no independence of A or b is required.

The last sum is bounded by the same sum over all keys with an observed unique majority. Conditional only on the key memberships, the labels within a key are independent Bernoulli(η_t) after integrating over their within-key features. Given $N_t = k > 0$, Hoeffding bounds the probability of an incorrect majority, including ties for an upper bound, by $e^{-k\gamma_t^2/2}$. Using the bound one when $k = 0$ and averaging the binomial count gives

$$\mathbb{E}\big[p_t\gamma_t \mathbf{1}\{N_t > 0,\ \hat{c}_t \neq c_t^*\}\big] \leq p_t\gamma_t(1 - p_t + p_t e^{-\gamma_t^2/2})^m \leq p_t\gamma_t e^{-mp_t\gamma_t^2/4}.$$

The last inequality uses $1 - e^{-u} \geq u/2$ for $0 \leq u \leq 1/2$. Summing proves (14). For the uniform bound, write each summand as $\sqrt{p_t/m}\, v_t e^{-v_t^2/4}$ with $v_t = \gamma_t\sqrt{mp_t}$. Then $\sup_{v \geq 0} v e^{-v^2/4} = \sqrt{2/e} < 1$ and Cauchy–Schwarz give $E_m(T) \leq \sqrt{J/m}$. If $\gamma_t \geq \gamma > 0$, the substitution $u = mp_t\gamma_t^2/4$ and $ue^{-u} \leq 1/e$ bound each summand by $4/(em\gamma)$.

When labels are deterministic given T, both $\Gamma_T(A)$ and the majority error term vanish for every fitting set. The identity therefore gives $R(h_A) \leq R(b)$. If every observed key is selected, errors can occur only on unseen keys, so $R(h_A) \leq U_D$. Finally, $\mathbb{E}U_D = \sum_t p_t \mathbb{P}(N_t = 0) = \sum_t p_t(1-p_t)^m$.

Accounting for minimum group support. If the gate selects only keys with $N_t \geq s$ for a fixed integer $s \geq 1$, the same proof replaces $E_m(T)$ by the finer quantity

$$E_{m,s}(T) = \sum_t p_t\gamma_t \sum_{k=s}^{m} \binom{m}{k} p_t^k (1-p_t)^{m-k} e^{-k\gamma_t^2/2}.$$

In particular,

$$E_{m,1}(T) = \sum_t p_t\gamma_t\big\{(1 - p_t + p_t e^{-\gamma_t^2/2})^m - (1-p_t)^m\big\}.$$

This removes unobserved keys, on which the rule retains its baseline. Every $E_{m,s}$ is bounded by the main text's $E_m(T)$, and also by $\sum_t p_t\gamma_t \mathbb{P}\{\mathrm{Bin}(m, p_t) \geq s\}$. Thus the estimation cost contracts with the region eligible to vote. The benefit $V_A(b)$ is restricted to the same selected region, so stronger support requirements also reduce the errors sharing can remove. No independence of the gate or baseline is added.

Decision sufficiency and localized approximation. If a Bayes decision is constant within every selected key, then $\Gamma_T(A) = 0$. Posterior equality is unnecessary. In general the absolute-value Jensen gap gives

$$\Gamma_T(A) \leq \Delta_A := \sum_{t \in A} p_t \mathbb{E}[|\eta_X - \eta_t| \mid T = t].$$

Define $K_A = \sum_{t \in A} p_t I(Y; X \mid T = t)$. Conditional Pinsker and Cauchy–Schwarz give

$$\Delta_A \le \sqrt{\mathbb{P}(T(X) \in A) K_A / 2} \le \sqrt{I(Y; X \mid T)/2}.$$

These inequalities hold for every realized A, including a data-dependent gate. They show why sufficient selected groups can be useful even when the partition discards label information elsewhere. They do not establish that the empirical purity gate finds such groups.

If the full posterior has a margin $|2\eta_X - 1| \ge \gamma > 0$ almost surely, the classification bias also satisfies

$$\Gamma_T(A) \le K_A / \gamma.$$

To prove this within one key, put $S = 2\eta_X - 1$ and change signs so that $\mu = \mathbb{E}[S \mid t] \ge 0$. Its classification bias is $b_t = \mathbb{E}[(-S)_+ \mid t]$. Conditional Pinsker gives $I(Y; X \mid t) \ge \mathrm{Var}(S \mid t)/2$. When $0 \le \mu \le \gamma$, the inequality $S^2 \ge \gamma |S|$ gives

$$\mathrm{Var}(S \mid t) \ge \gamma(\mu + 2b_t) - \mu^2 \ge 2\gamma b_t.$$

When $\mu \ge \gamma$, restrict the variance integral to $S < 0$ and use $(|S| + \mu)^2 \ge 4\mu|S|$ to obtain $\mathrm{Var}(S \mid t) \ge 4\mu b_t \ge 2\gamma b_t$. Hence $I(Y; X \mid t) \ge \gamma b_t$. Weight by p_t and sum over A. This sharpens the information penalty for the Bayes classification gap; it does not replace the arbitrary-comparator penalty in Theorem 3.1.

A.4 Selecting Explanation Maps from Data

The fixed-key results extend to a described family by standard simultaneous bounds. This section states the selection cost explicitly; it does not treat unrestricted pattern discovery as free.

Fix a countable family $(T_j)_{j \in \mathcal{J}}$ and positive masses π_j with $\sum_j \pi_j \le 1$, before drawing the fitting sample. Let $d_j = \mathrm{VC}_{T_j}(\mathcal{H})$ and $\Delta_j = \mathbb{E}|\eta_X - \eta_{T_j(X)}|$. The family may include every finite pattern set in a chosen countable description language. A prefix-free description of length L_j bits permits $\pi_j = 2^{-L_j}$ by Kraft's inequality.

Corollary A.16 (Simultaneous pooling guarantee)**.** *With probability at least $1 - \delta$, every finite-dimension map T_j and every score $q_j = Q_D^{T_j} h_j$ fitted to objective tolerance ξ_j satisfy*

$$R(q_j) \le R^*_{\mathcal{H}} + 3\Delta_j + \xi_j + 4C_0 \sqrt{\frac{d_j + 1}{m}} + \frac{7}{\sqrt{2}} \sqrt{\frac{\log(4/(\delta \pi_j))}{m}}. \tag{24}$$

The map, fitted hypothesis, and tolerance may be selected using the fitting data.

Proof. For a fixed map, the proof of Theorem A.5 is uniform in the fitted hypothesis after adding its empirical objective gap. In more detail, define

$$S_j(D) = \sup_{h \in \mathcal{H}} \left[R(Q_D^{T_j} h) - \left\{ \widehat{R}_D(Q_D^{T_j} h) - \inf_g \widehat{R}_D(Q_D^{T_j} g) \right\} \right].$$

The same feature and label events bound this quantity using one fixed comparator. If the population infimum is unattained, the near-minimizer argument applies to this single random variable S_j by continuity of increasing exceedance events. Thus the good event holds for all empirical objective gaps simultaneously. Apply it with failure probability $\delta \pi_j$ and take a countable union bound. The event then permits any data-dependent choice from the fixed ambient family. □

A mined shortlist can depend on the data. Its size alone is not the fitting complexity: the bound charges its members' ambient prior masses or description lengths. An independent earlier development sample is another way to fix a map before fitting, in which case no family selection cost is needed.

Validation without estimating sufficiency. Let a finite nonempty shortlist $\mathcal{J}_D$ be proposed using fitting data, with $N = |\mathcal{J}_D|$. Fit its scores on D and, optionally, a binary baseline b_D. All these functions must be fixed before a fresh validation sample of size v is drawn. Select the smallest validation error over the baseline and all thresholds $\mathbf{1}\{q_j \ge u\}$, $j \in \mathcal{J}_D$, $0 \le u \le 2$.

Conditional on D, the threshold argument used above costs $4e^{-2va^2}$ per score family, $2e^{-2va^2}$ for the common class prior, and $2e^{-2va^2}$ for the baseline. Thus uniform validation deviation is at most $3a$ with failure at most $(4N+4)e^{-2va^2}$. Combining it with Corollary A.16 at confidence $\delta/2$ gives

$$R(\hat{g}) \le \min\left\{R(b_D), \min_{j\in\mathcal{J}_D}\left[R^*_{\mathcal{H}} + 3\Delta_j + \xi_j + 4C_0\sqrt{\frac{d_j+1}{m}} + \frac{7}{\sqrt{2}}\sqrt{\frac{\log(8/(\delta\pi_j))}{m}}\right]\right\} + 6\sqrt{\frac{\log((8N+8)/\delta)}{2v}} \tag{25}$$

with probability at least $1-\delta$. The threshold integral ensures $\inf_u R(\mathbf{1}\{q_j \ge u\}) \le R(q_j)$. The shortlist size may be random because it is fixed conditional on D. If the baseline is omitted, remove its entry in the minimum; the displayed constants remain valid. This selects using validation error without requiring knowledge of any Δ_j or κ_j.

Preserving a fast rate during selection. A generic validation bound would add a $v^{-1/2}$ term to a fast-rate candidate. Under the global margin condition $|2\eta_X - 1| \ge \gamma > 0$, finite-candidate validation has a sharper bound. Conditional on fitting data, let $g_1, \dots, g_N$ be any fixed binary candidates, such as the half-thresholds of pooled scores, and let $\hat{g}$ minimize validation error to tolerance ν. Then

$$R(\hat{g}) - R^* \le 4\min_j\{R(g_j) - R^*\} + 2\nu + \frac{6\log(2N/\delta)}{\gamma v} \tag{26}$$

with probability at least $1-\delta$. Neither the Bayes rule nor γ is used by the selector.

To prove this, fix a Bayes rule h^* and put $E_j = R(g_j) - R^*$ and $\hat{E}_j = \widehat{R}_V(g_j) - \widehat{R}_V(h^*)$. The one-example excess $W_j = \ell(g_j(X), Y) - \ell(h^*(X), Y)$ lies in $\{-1, 0, 1\}$, with $\mathbb{E}W_j = E_j$ and $\mathbb{E}W_j^2 \le E_j/\gamma$. Using $e^u \le 1 + u + u^2$ for $|u| \le 1$, independence gives

$$\mathbb{E}e^{(\gamma v/4)(E_j - 2\hat{E}_j)} \le 1, \qquad \mathbb{E}e^{\gamma v(\hat{E}_j - 2E_j)} \le 1.$$

For the first bound, use $u = -\gamma W_j/2$; for the second, use $u = \gamma W_j$. Markov and a union bound give $E_j \le 2\hat{E}_j + 4\log(2N/\delta)/(\gamma v)$ for every candidate, except with probability $\delta/2$. The second moment bound gives $\hat{E}_{j^*} \le 2E_{j^*} + \log(2/\delta)/(\gamma v)$ for the population-best candidate j^*, except with probability $\delta/2$. It is fixed conditional on fitting data. Empirical minimization now proves (26); the unknown Bayes empirical risk cancels from the selection rule.

The fast pooling theorem also admits the prior-weighted union bound. Consequently, if the shortlist contains an exactly sufficient map and the Bayes classifier belongs to $\mathcal{H}$, splitting the total sample into fitting and validation sets of comparable size preserves a fast rate, with a description cost during fitting and a $\log N$ cost during selection. Decision-sufficient candidates can instead use the γ^{-2} fitting bound. These statements compare Bayes excess multiplicatively; they are not additive guarantees of improvement over an arbitrary baseline.

Adaptive keys for selective majority votes. For a fixed finite map with J keys, a proof-only histogram predictor can use empirical majority labels on observed keys and the true Bayes-T label elsewhere. Resolve observed ties in favor of that Bayes label. This predictor is an empirical minimizer over the 2^J groupwise binary classifiers. Uniform Hoeffding implies, with probability at least $1-\delta$,

$$\sum_{t\in A} p_t\gamma_t \mathbf{1}\{\hat{c}_t \ne c^*_t\} \le 2\sqrt{\frac{J\log 2 + \log(2/\delta)}{2m}}$$

simultaneously for every gate selecting observed keys with a unique majority. The selected error is bounded by the histogram's excess over its Bayes-T risk. Combining this event with the pointwise identity (23) covers every baseline and gate.

For the countable ambient family, replace J by J_j and δ by $\delta\pi_j$, and take a union bound. This permits data-dependent map selection, including additional development data, while charging the ambient family. It does not give a small numerical certificate automatically: K selected binary patterns can induce as many as 2^K groups. The original graph studies also use stratified samples, rather than the i.i.d. protocol assumed here. These distinctions remain part of the interpretation of the experiments.

B Numerical Protocol and Verification

The numerical study stores a loss-equivalent graph encoding as an explanation key and a background identifier. Empirical objectives are optimized exactly. Population risk uses the randomized loss (1) and is computed analytically, with no test-label or prediction-randomness Monte Carlo. Every method receives the same m i.i.d. labeled graphs.

Distribution and encoding. Key zero has probability 0.85 and a background uniform on $\{0, \dots, N_{\mathrm{bg}} - 1\}$. Each of the J tail keys has probability $0.15/J$. In key zero, $Y \sim \mathrm{Bernoulli}(\eta_0)$. On tail key t, Y equals $\mathbf{1}\{t > J/2\}$ with probability 0.9. Every reported J is even. Labels depend only on the key, so $\kappa = 0$. An attributed graph can encode the key by a marked cycle and the background by an attached rooted attributed tree. The simulation stores the two identifiers directly; the tree encoding does not affect any risk or objective.

Hypothesis class. For every cutoff $c \in \{0, \dots, J\}$, include the classifier that predicts zero on key zero and $\mathbf{1}\{t > c\}$ on the tail. Also include all classifiers that are arbitrary functions of the N_{bg} recurrent backgrounds and predict one on the entire tail. The Bayes classifier belongs to the threshold family, with risk $0.85\eta_0 + 0.15(0.1)$.

The class shatters all N_{bg} recurrent backgrounds. It cannot shatter a set containing both a recurrent point and a tail point: label one on the former forces label one on every tail point. The tail threshold family has VC dimension one. Thus ordinary VC dimension is N_{bg}, and across-key VC dimension is one. A finite verifier checks the latter by enumeration for $N_{\mathrm{bg}} = 1, 2, 4, 6, 8$.

Exact fitting. Ordinary ERM compares the best tail threshold with the best background memorizer. A threshold minimizes cumulative tail error as its cutoff passes observed keys. The background branch minimizes error separately at every observed background, assigning zero to unseen or tied backgrounds. Ties favor the threshold branch and then its smallest cutoff.

For pooled ERM, a represented recurrent key has a constant score. The background branch contains both endpoint scores, zero and one. Linearity of the fitting loss means an endpoint is optimal in that branch. Pooled ERM therefore compares the best threshold with these two background classifiers. If the recurrent key is absent, the selected hypothesis is retained there; the threshold branch wins ties and is at least as good as the all-one tail. Tail predictions are already constant by key. These cases implement exactly (6) and (7). Integer loss numerators preserve the specified tie rule.

The majority-sharing comparison applies each unique fitting-group majority to the ordinary ERM predictor, retaining it on ties and unseen keys. The random-label comparison predicts one with the empirical fitting-label frequency in every seen key, retaining the same fallback elsewhere. Risks include covered and uncovered tail keys. All comparisons receive exactly the same labels and features.

Configurations. The noise study fixes $N_{\mathrm{bg}} = 1{,}048{,}576$, $J = 5{,}000$, $m \in \{5, 10, 20, 50, 100, 200, 500, 1000\}$, and $\eta_0 \in \{0.05, 0.20, 0.35, 0.45\}$. Each of its 32 cells uses 3,000 independent training repetitions. Capacity studies use $\eta_0 = 0.35$ and $m \in \{10, 50, 200, 1000\}$. One varies $N_{\mathrm{bg}} \in \{8, 32, 128, 1024, 1{,}048{,}576\}$ at $J = 5{,}000$; the other varies $J \in \{4, 20, 100, 1000, 10{,}000\}$ at $N_{\mathrm{bg}} = 1{,}048{,}576$. Each of these 40 cells uses 2,000 repetitions. The seed is 2026092508. Figure 1(b,c) shows the $m = 200$ slice; accompanying results retain every cell. Error bars describe training-sample variation in exact population risk.

B.1 Non-Invariant Comparators and Approximate Sufficiency

Modify the threshold branch so that its recurrent prediction is the fixed background bit $v(b) = \mathbf{1}\{b < N_{\mathrm{bg}}\theta\}$, with $N_{\mathrm{bg}} = 10^6$ and $\theta = 0.1$. The rest of the class is unchanged. Across-class VC dimension remains one: for a recurrent point with $v = 0$, labels $(1, 0)$ on it and a tail point are unattainable; for $v = 1$, labels $(0, 0)$ are unattainable. Two tail points cannot be shattered. Ordinary VC dimension remains N_{bg}.

Set $\eta_0 = 0.35$ and let the recurrent label probability be

$$\mathbb{P}(Y = 1 \mid b) = \begin{cases} \eta_0 + a(1 - \theta), & v(b) = 1, \\ \eta_0 - a\theta, & v(b) = 0. \end{cases}$$

The recurrent mean remains η_0. At $a = 0$ the key is sufficient. For every tested a, the best hypothesis is the true tail threshold with recurrent bit v, whose risk is

$$R^*_{\mathcal{H}} = 0.85\{\eta_0 + \theta(1 - 2\eta_0) - 2a\theta(1 - \theta)\} + 0.15(0.1).$$

At $a = 0$ this risk is 0.338, whereas the best invariant member of $\mathcal{H}$ has risk 0.3725: all invariant members predict one on the entire tail. The comparator's population projection cost is $\Delta_T = 2(0.85)a\theta(1 - \theta)$. Conditional information is computed from the two Bernoulli divergences to their mixture mean η_0.

The pooled threshold branch uses the empirical average $\bar{v}$ on recurrent training graphs. Its recurrent fitting loss is $A + (k - 2A)\bar{v}$, where k is the recurrent count and A its positive-label count. The background branch still has feasible endpoint predictions. Multiplying these losses by k gives integer numerators for exact ERM, including ties. The average $\bar{v}$ and the labels come from the same sample and can be dependent; the implementation preserves this dependence.

The comparator grid uses $m \in \{5, 10, 20, 50, 100, 200, 500, 1000, 2000\}$ at $a = 0$, with $4{,}000$ repetitions per cell. The information-loss grid uses $a \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and $m \in \{50, 200, 1000\}$, with $3{,}000$ repetitions per cell. They use seed 2026092508 with disjoint offsets from the first construction. Figure 2 displays both complete grids. The four studies contain 108 configurations and $293{,}000$ independent fitting samples.

Independent implementation checks. Verification scripts construct literal finite hypothesis matrices rather than using the study's optimizer formulas. Across $5{,}000$ cases with varied sample sizes, key counts, background bits, and conditional label probabilities, their population risks match the study's actual implementation to 3.34×10^{-16}. The oracle formula is separately checked at all nine information-loss parameters, to 1.67×10^{-16}.

For the proof, complete enumeration of $7{,}060$ balanced allocations and $35{,}202$ representative choices verifies representative/count independence on 54 pooled count profiles through sample size five, with errors below 3.34×10^{-16}. Another check enumerates $56{,}160$ labeled samples from twelve finite distributions, with sixteen hypotheses per sample: the mean comparator covariance identity agrees to 7.41×10^{-15}, and the soft-target and weighted-majority objective identities to 1.12×10^{-16}. An additional $182{,}104$ group-count states verify the addition bound used in comparator concentration. The risk-dependent proof is also checked under unequal posteriors within keys: $635{,}375$ group-count states satisfy the proved squared-deletion bound, and 1600 fixed-comparator moment comparisons satisfy its exponential bound. Separate exact sums check the label offsets, relative threshold validation, and conditional-averaging variance identity. These finite checks supplement the proofs; distribution-uniform validity rests on Appendix A.

B.2 Larger Budgets and the Fast Rate

We extend the exact-sufficiency construction with $J = 5000$ to $m \in \{20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000\}$ and recurrent-class noise $\eta_0 \in \{0.20, 0.35, 0.45\}$. Each of the 30 configurations uses 3000 independently sampled fitting sets. The Bayes classifier belongs to $\mathcal{H}$ and the minimum conditional margin is $\min\{0.8, 1 - 2\eta_0\}$, so Theorem 3.2 applies.

For efficiency, this experiment draws the multinomial counts of recurrent-class labels and each tail key/label pair. These counts are sufficient for the same exact pooled objective. A tail threshold minimizes the prefix sums of $n_{t,1} - n_{t,0}$, and the competing arbitrary-background branch attains its minimum at a constant recurrent-class endpoint. Integer losses and the same threshold-first tie convention avoid numerical ambiguity. We checked this count-based optimizer against all explicit candidates in 1500 small instances: the objectives agree exactly, and the population-risk formulas differ by at most 5.6×10^{-17}.

Figure 3 shows a regime in which multiplying mean excess risk by m gives an approximately stable scale. At $m = 20000$, mean excess is about 5×10^{-5}. The guide is descriptive and does not estimate a universal theorem constant or establish a rate by regression. Because the tail has finitely many keys, still larger samples can yield faster decay. The figure contains all configurations; bands show 1.96 Monte Carlo standard errors, with population risk evaluated exactly for each fitted predictor. No reference or validation sample is used. The selected head endpoints and tail thresholds are binary, so half-thresholding gives the same risk.

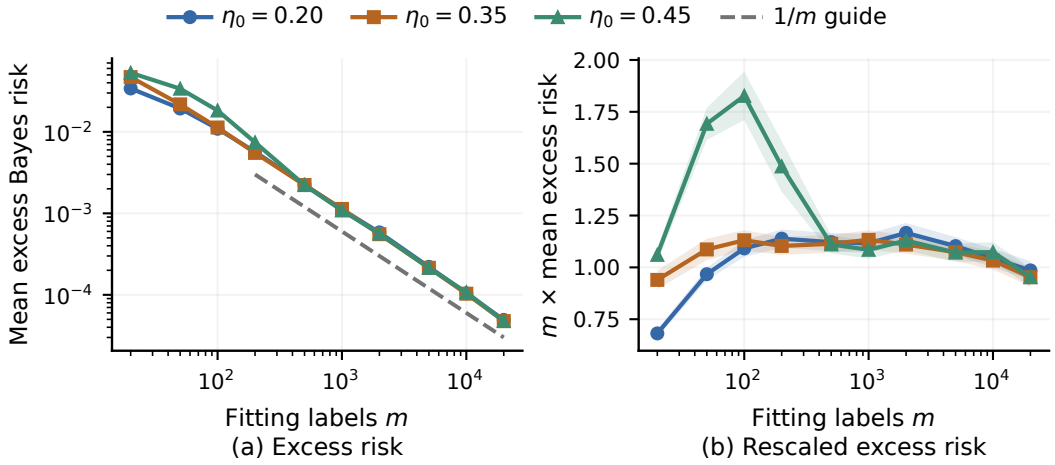


Figure 3: The same pooled ERM at larger fitting budgets under Bayes realizability and a positive conditional margin. Left: mean excess Bayes risk at recurrent-class noise η_0; the dashed line is a $1/m$ guide. Right: the same means multiplied by m. Each point uses 3000 fitting samples; bands show 1.96 Monte Carlo standard errors.

Varying the conditional margin. A second sweep sets the label-flip probability to $(1 - \gamma)/2$ in every class, for $\gamma \in \{0.05, 0.1, 0.2, 0.4, 0.8\}$. It uses the same hypothesis family, recurrent mass 0.85, and $J = 10^6$ tail keys. Budgets are $m \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000, 100000\}$, with 3000 fitting samples per cell and seed 2026092518. Sparse labeled counts implement the identical pooled ERM; 3000 additional literal candidate comparisons verify its integer objectives and tie choices exactly.

All 55 configurations appear in Figure 4. Plotting excess risk divided by γ against $m\gamma^2$ aligns the curves. In the larger-sample regime, $m\gamma$ times mean excess stays near a constant for each margin. This tests the $1/(m\gamma)$ dependence in Theorem 3.2, beyond varying the recurrent class's noise alone. The rescaling is descriptive; it does not establish sharp constants or optimality. The complete numerical record, including the dimension and small-risk studies below, contains 862,000 fitting repetitions, with all configurations exported.

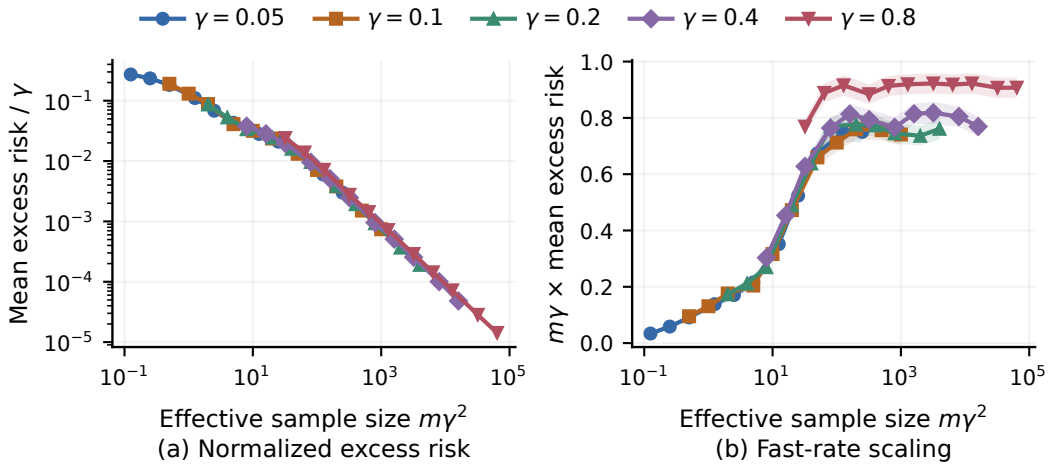


Figure 4: Conditional-margin dependence for the same pooled ERM. Every class has margin γ, and the Bayes classifier belongs to $\mathcal{H}$. Left: normalized excess against the effective sample size. Right: the same results on the scale suggested by the fast-rate bound. Bands show 1.96 Monte Carlo standard errors over 3000 fitting samples per configuration.

Polynomial margin tails. A third study tests Proposition A.9 using the same recurrent-class/tail hypothesis family. The recurrent class keeps mass 0.85 and noise 0.35. On tail key $t \in \{1, \dots, J\}$, with $J = 10^6$, set

$$\gamma_t = \left| \frac{2t - J - 1}{J} \right|^{1/\alpha}, \qquad \mathbb{P}(Y \neq \mathbf{1}\{t > J/2\} \mid T = t) = \frac{1 - \gamma_t}{2}.$$

We fix $\alpha \in \{0.5, 1, 2\}$ and $m \in \{20, 50, 100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000\}$, with 3000 iid fitting repetitions per cell and seed 2026092526. The protocol is saved before evaluation and all 33 configurations are retained. The Bayes threshold remains in $\mathcal{H}$ and $d = 1$. The discrete tail obeys $\mathbb{P}(\gamma_T \leq u \mid T \neq 0) \leq 2u^\alpha$, so (19) holds with $A = 0.3 + 0.85/0.3^\alpha$, independently of J.

Joint head-label counts and sparse tail observations implement the same exact objective. These statistics are sufficient regardless of the recurrent class's background capacity. For a selected head endpoint q and tail cutoff c, its exact excess is

$$0.85(0.3)q + \frac{0.15}{J} \left| \sum_{t=1}^{c} \gamma_t - \sum_{t=1}^{J/2} \gamma_t \right|.$$

Complete finite-hypothesis enumeration checks the optimizer in 2000 instances, with risk differences below 7×10^{-18}. The margin-tail label offset is separately checked by 1200 complete label sums, including zero and near-zero margins; 2000 exact moment checks cover concentration of its remainder.

Figure 5 plots all cells. Multiplying mean excess by $m^{(\alpha+1)/(\alpha+2)}$ gives approximately stable scales at larger budgets, consistent with the margin-tail rate. The finite grid still has a positive smallest margin, so this tests an intermediate regime rather than a limiting rate at fixed J. The exponents are set by the protocol, not fitted to these curves. No validation sample is used; the fitted scores are binary.

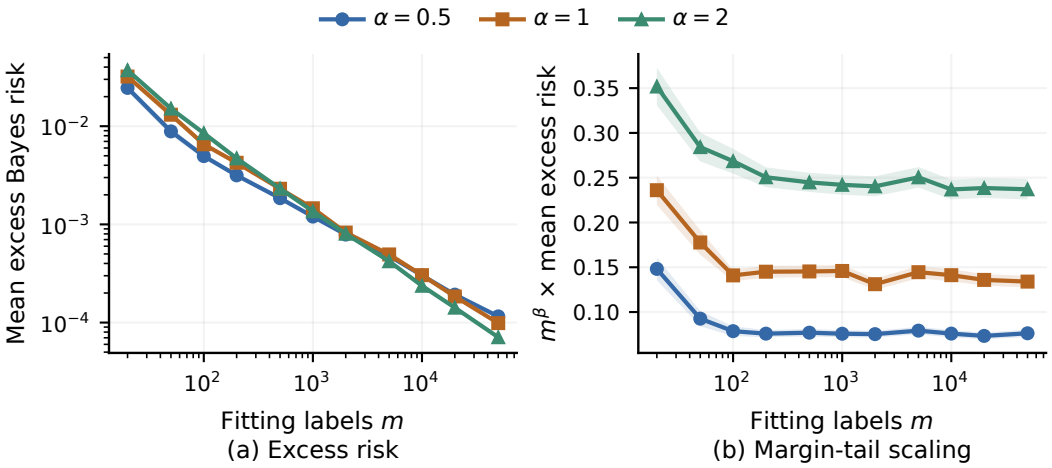


Figure 5: The same pooled learner with polynomial mass near the label boundary. Left: exact population excess, averaged over 3000 fitting samples. Right: rescaling by m^β, where $\beta = (\alpha + 1)/(\alpha + 2)$ is specified by each margin distribution. Bands show 1.96 Monte Carlo standard errors. All 33 configurations are shown.

B.3 Varying Across-Class Dimension

We take the direct product of $d \in \{1, 2, 4, 8, 16\}$ copies of the verified recurrent-class/tail construction. Each graph includes an observed block identifier as part of its structural key. Blocks have equal probability; their local distributions use $N_{\text{bg}} = 1048576$, $J = 5000$, and recurrent-class noise $\eta_0 = 0.35$. A hypothesis chooses its local classifier independently in each block. The across-class VC dimension is exactly d: one recurrent-class point per block can be shattered, whereas any larger

set with distinct keys contains two points from some block, whose local across-class dimension is one. The ordinary VC dimension is dN_{bg}.

Every fitting sample contains exactly m i.i.d. labeled observations. Their block counts are multinomial, rather than fixed equal quotas. The empirical objective separates across blocks, so the simulation calls the actual previously verified optimizer on each local sample and averages their exact population risks. Empty blocks use the same fixed tie convention. We use $m \in \{20, 50, 100, 200, 500, 1000, 2000, 5000\}$ and 1000 complete fitting samples per cell, retaining all 40 configurations and all four methods. Figure 6 shows the pooled learner before and after rescaling m by d. This construction tests the dimension dependence in the Bayes-realizable margin regime; it is not an estimate of d for a trained GNN.

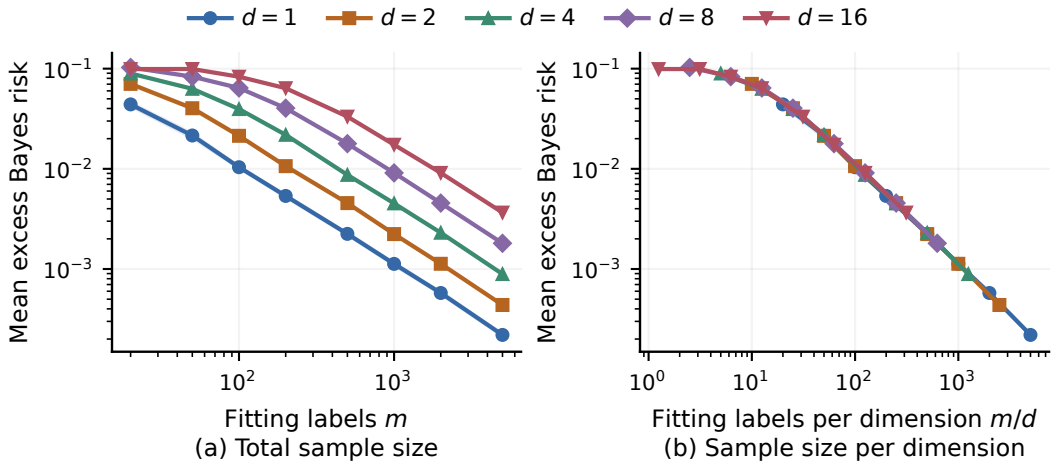


Figure 6: Across-class complexity in the product construction. Left: increasing d raises the sample requirement. Right: the same complete grid plotted against m/d. Bands show 1.96 Monte Carlo standard errors across 1000 full fitting samples.

B.4 Small Comparator Risk under Misspecification

We vary the comparator risk while keeping $d = 1$ and $\Delta = 0$. Use the non-invariant class from Appendix B.1, with $N_{\mathrm{bg}} = J = 10^6$ and recurrent mass 0.85. Set the recurrent label to zero and make tail labels deterministic at the true cutoff $J/2$. The threshold branch retains its fixed recurrent prediction $v(b) = \mathbf{1}\{b < N_{\mathrm{bg}}\theta\}$. We use $\theta \in \{0.001, 0.003, 0.01, 0.03, 0.06\}$, giving $R^*_{\mathcal{H}} = 0.85\theta$. The best arbitrary-background branch has risk 0.075, above every tested $R^*_{\mathcal{H}}$. The Bayes classifier is absent from $\mathcal{H}$, so the Bayes-realizable margin theorem does not apply.

If the fitting sample contains K recurrent graphs, V of them with $v = 1$, and N_- negative tail graphs, the threshold and background branches have optimal empirical error counts V and N_-. The same pooled ERM compares these integers, favoring the threshold branch on ties. Its smallest zero-error tail cutoff is the largest observed negative key. Counts are drawn jointly from their multinomial law. Conditional on $N_- = n > 0$, this largest key has distribution $\lceil (J/2)U^{1/n} \rceil$, with U uniform on $(0, 1)$; it is zero when $n = 0$. This generates exactly the statistics used by the learner from m iid observations. When the threshold branch is selected, its recurrent pooled score is V/K for $K > 0$, and its population tail error is known from the cutoff.

The protocol fixes $m \in \{100, 300, 1000, 3000, 10000, 30000, 100000\}$, 5000 fitting repetitions per cell, and seed 2026092522 before evaluation. All 35 configurations appear in Figure 7. We report the 95th percentile of $R(\hat{q}) - R^*_{\mathcal{H}}$, since Proposition A.12 controls upper deviations. Its rescaled value is about 1.62–1.81 at $m = 100000$, consistent with the $\sqrt{R^*_{\mathcal{H}}/m}$ term. The full record also retains signed mean risk, positive excess, and branch frequencies. For the threshold candidate before branch selection, the recurrent risk contribution has conditional variance $0.85^2\theta(1-\theta)/K$; this explains why averaging signed excess alone can hide the variation of interest. As in the other numerical studies, the evaluated output is the pooled score, without threshold-validation labels.

Literal finite-hypothesis enumeration checks the count implementation on 2000 instances: objective differences are below 2.67×10^{-15} and population-risk differences below 1.12×10^{-16}. Exact finite sums check the conditional maximum distribution in 126 cases. The percentile intervals use binomial order-statistic ranks 4719 and 4781 among 5000 sorted repetitions, giving at least 95% coverage. This experiment tests the refinement's dependence on comparator risk; it does not establish an optimal learning rate.

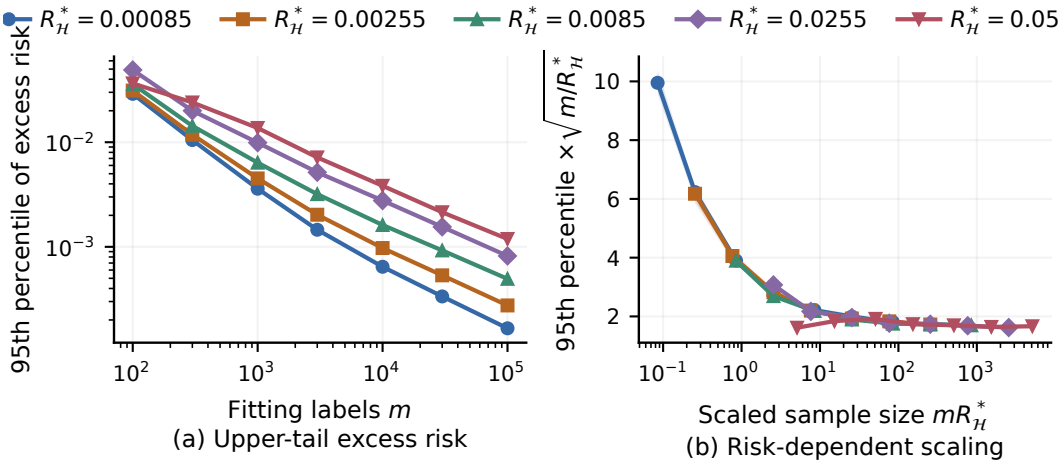


Figure 7: Risk-dependent learning with a misspecified hypothesis class. The Bayes risk is zero but $R^*_{\mathcal{H}} > 0$. Left: the 95th percentile of the pooled score's excess over $R^*_{\mathcal{H}}$. Right: the complete grid rescaled by $\sqrt{R^*_{\mathcal{H}}/m}$. Bands are 95% order-statistic confidence intervals for the percentile across 5000 fitting samples per cell. Population risks are exact.

C Experimental Setup and Protocols

C.1 Datasets and Construction

We use the datasets introduced in Section 5 for three experimental purposes. BA2MOTIFS and the synthetic explanation-class family provide controlled tests; MUTAGENICITY evaluates supplied explanations; and FOUR-MOTIF, AIDS, TOX21, DHFR, and PROTEINS evaluate the mined application, with BA2MOTIFS also providing a structural baseline.

The molecular and protein collections are distributed through TUDataset (Morris et al., 2020). Table 6 summarizes the graph counts, class proportions, attributes, and splits used in our experiments.

BA2MOTIFS. BA2MOTIFS contains 1,000 graphs with constant node attributes and balanced binary labels (Luo et al., 2020). Each graph combines a twenty-node Barabási-Albert background (Barabási & Albert, 1999) with a five-node house or cycle motif attached by one edge. The motif supplies the label and explanation; Appendix D.1 compares ordinary and induced matching.

Synthetic explanation-class family. The controlled family has five datasets of 4,337 graphs, with $L \in \{2, 4, 6, 8, 10\}$ available motif types. Labels depend on the planted motif set, while graph size and node-type counts are controlled to prevent simple shortcuts. The construction below fixes the background statistics; Appendix D.1 analyzes the resulting label requirements.

MUTAGENICITY. MUTAGENICITY contains 4,337 molecular graphs with binary mutagenicity labels (Kazius et al., 2005). We attach ten fixed structural descriptors that approximate published toxicophore patterns using atom types and adjacency. These descriptors are defined before fitting and provide the supplied explanations used in Section D.2. Appendices C.4–C.6 give extraction, voting, and controls.

FOUR-MOTIF. FOUR-MOTIF is a synthetic task of 4,337 graphs designed to require combinations of structural patterns. Each graph carries $k \in \{0, 1, 2, 3, 4\}$ motifs with probabilities 0.15, 0.35, 0.30,

Table 6: Dataset statistics for the experiments in Section 5. Balance gives the two class proportions; Types counts distinct node-attribute values. $\overline{|V|}$ and $\overline{|E|}$ are mean node and undirected edge counts. Fit/Val/Test gives the reference stratified 80/10/10 partition with seed 12345. The synthetic family contains five datasets, one for each motif vocabulary size, and its node and edge means are averages over the five.

Dataset	Kind	Graphs	Balance	Types	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	Fit/Val/Test
BA2MOTIFS	synthetic	1,000	.500/.500	1	25.0	25.5	800/100/100
Synthetic family	synthetic family	4,337 each	≈.500/.500	8	≈50.1	≈57.1	3,470/434/433
MUTAGENICITY	annotated	4,337	.554/.446	14	30.3	30.8	3,470/434/433
FOUR-MOTIF	synthetic	4,337	.499/.501	8	49.9	56.9	3,470/434/433
AIDS	molecules	2,000	.200/.800	38	15.7	16.2	1,600/200/200
TOX21	derived	1,900	.500/.500	28	18.6	19.3	1,520/190/190
DHFR	molecules	756	.390/.610	9	42.4	44.5	605/76/75
PROTEINS	proteins	1,113	.596/.404	3	39.1	72.8	890/111/112

0.15, and 0.05, and its label indicates whether $k \geq 2$. The sixteen motif subsets define the population explanation groups and determine their labels without noise. The realized counts for $k = 0, \ldots, 4$ are 616, 1,559, 1,288, 663, and 211; graphs carrying no motif form the empty-key group. Node and edge totals and node-type histograms are drawn before planting motifs to prevent these counts from revealing the label. Class means are 50.0 and 49.9 nodes, with 56.9 edges for both.

AIDS. AIDS is a molecular activity dataset of $2{,}000$ graphs from the IAM repository (Riesen & Bunke, 2008), with class proportions 0.200 and 0.800. The graphs average 15.7 nodes, and graph size is associated with the label, motivating the graph-size control defined in Appendix C.6.

TOX21. For TOX21, we use a balanced sample of $1{,}900$ molecules from the AhR assay distributed through TUDataset (Morris et al., 2020). We retain all 950 minority-class molecules from the $8{,}169$-graph source collection and sample 950 majority-class molecules without replacement using seed 12345, then form a stratified split of this balanced sample. The TOX21 results throughout the paper refer to this preprocessing and class distribution.

DHFR. DHFR has 756 molecular graphs from activity data (Sutherland et al., 2003), with class proportions 0.390 and 0.610, mean size 42.4 nodes, and 75 test graphs.

PROTEINS. PROTEINS contains $1{,}113$ protein contact graphs with node attributes (Dobson & Doig, 2003; Borgwardt et al., 2005). The graphs average 39.1 nodes, with class proportions 0.596 and 0.404. All PROTEINS experiments use this same attributed collection.

Construction of the explanation-class family. Each dataset contains $4{,}337$ graphs generated with seed 901 from the first L motifs in a fixed ten-motif vocabulary. For each graph, the number R of planted motifs is drawn from $\{0, 1, 2\}$ with probabilities 0.15, 0.35, and 0.50; conditional on R, the generator selects R distinct motifs uniformly without replacement. The label is $Y = 0$ when $R = 2$ and $Y = 1$ otherwise, making the population classes balanced and the label a deterministic function of the planted motif set. That set is the population explanation key. Consequently, the possible keys are the empty set, the L singleton sets, and the $\binom{L}{2}$ pairs, so

$$C_L = 1 + L + \binom{L}{2}. \tag{27}$$

All possible keys occur in the realized datasets, giving $C_L = 4, 11, 22, 37, 56$ as L increases.

The graph count, distribution of R, label rule, node-type alphabet, and random seed are fixed across the family. The generator draws the target node and edge totals before planting the motifs and adjusts the background graph to absorb their sizes, so those totals do not reveal R. The motif node types are drawn from the same eight-type alphabet under a fixed balanced construction.

C.2 Splits, Fitting Budgets, and Label Access

The reference partition uses a stratified 80/10/10 split into fitting, validation, and test graphs, with seed 12345. The test split is used only for final evaluation, after mining, pattern selection, voting, and checkpoint selection. Three-split curves measure partition sensitivity on the same graphs.

The fitting budget t_n takes the values 2, 5, 10, 20, 50, 100, 200, and Full, counted per class, with Full denoting the complete fitting pool. For each seed the fitting subset at budget t_n is the prefix of a classwise random permutation, so the subset at one budget is contained in the subset at the next. Curve points are therefore not independent draws: moving right along a curve keeps the same graphs and adds more. Seeds 42–51 give 30 runs per point across three splits and 10 per cell for the original-split matching and MUTAGENICITY studies.

The theorems assume independent sampling from a fixed distribution. Class-balanced nested sampling tests the predicted mechanism under a different protocol; its results do not estimate theorem constants.

Selection samples at most $2t_n$ validation graphs uniformly without replacement, using all validation graphs at Full. The label budget counts the union of indices accessed across all roles:

$$N_{\text{labels}} = |I_{\text{fit}} \cup I_{\text{checkpoint}} \cup I_{\text{selection}} \cup I_{\text{vote}} \cup I_{\text{calibration}}|.$$

In this protocol, $I_{\text{vote}} = I_{\text{fit}}$ and $I_{\text{selection}} \subseteq I_{\text{checkpoint}} = I_{\text{val}}$, giving $N_{\text{labels}} = |I_{\text{fit}}| + |I_{\text{val}}|$ for both methods. For TOX21, $t_n = 200$ uses $400 + 190 = 590$ labels and Full uses $1{,}520 + 190 = 1{,}710$. For FOUR-MOTIF, $t_n = 50$ and 200 use $100 + 434 = 534$ and $400 + 434 = 834$ labels. Mining reads fitting-graph structure; supplied motifs provide additional structural information. The MUTAGENICITY experiments use the reference partition of 3,470 fitting-pool, 434 validation, and 433 test graphs. BA2MOTIFS matching and both MUTAGENICITY studies use GCN and GIN on the original split; the main application uses all four backbones on three splits.

C.3 Backbones and Implementation

The base classifiers are GCN (Kipf & Welling, 2017), GIN (Xu et al., 2019), GraphSAGE (Hamilton et al., 2017), and PNA (Corso et al., 2020), implemented in PyTorch Geometric (Fey & Lenssen, 2019). Each classifier has three message-passing layers and hidden width 64; GIN uses sum pooling and the other classifiers use mean pooling. PNA computes its degree histogram from the fitting graphs available in the corresponding run. All classifiers are trained for 200 epochs with AdamW (Loshchilov & Hutter, 2019), learning rate 0.005, zero weight decay, and batch size 256. Checkpoint selection keeps the first epoch attaining the highest accuracy on the full validation split and reloads those parameters before the single test evaluation. The rule uses frozen predictions without retraining, retaining the GNN output or replacing it with an eligible group vote.

The main curve experiments used four NVIDIA RTX 4080 GPUs, Python 3.12, NumPy 2.3.2, NetworkX 3.5, PyTorch 2.8.0 with CUDA 12.8, and PyTorch Geometric 2.8.0.post1. Pattern construction and rule evaluation run on CPUs, with the trained GNNs fixed. The pooled-ERM, structural-baseline, and strict-budget tuning studies of Appendices D.6 and E used PyTorch 1.12.1 and PyTorch Geometric 2.3.

C.4 Explanation Construction and Matching

Supplied explanations. BA2MOTIFS and FOUR-MOTIF supply their planted motifs. For MUTAGENICITY, the explanation vocabulary is fixed before fitting and extracted from graph structure. The supplied vocabulary contains ten descriptors: nitro, nitroso, N–O, amine, ring amine, N–N, aliphatic halide, acyl halide, a three-membered ring containing a heteroatom, and a polycyclic structure. They approximate published toxicophore patterns using atom types and adjacency (Kazius et al., 2005). The dataset lacks bond orders and aromaticity flags, so these predicates are approximate structural descriptors, not verified complete chemical explanations. The explanation key is the set of descriptor names matched in a graph; repeated occurrences of a descriptor do not change the key. The empty set is a valid key. The vocabulary and extraction rule do not use the fitting labels; only the group votes use them. The polycyclic descriptor depends on a cycle basis; the extraction procedure is held fixed within a comparison.

Mined explanations. All three miners use node attributes and adjacency deterministically.

`W` emits Weisfeiler-Leman subtree patterns (Shervashidze et al., 2011): a pattern is a pair of a depth in $\{1, 2, 3\}$ and a refinement color at that depth, and a graph carries it when some node attains that color. A depth-d color encodes the rooted unfolding after d refinement rounds; this refinement bounds the distinguishing power of standard d-layer message-passing GNNs (Xu et al., 2019).

Table 7: Label access in the shared-validation protocol: Fit (fitting subset), Val (validation), Test (evaluation).

Stage	GNN	Hybrid
Fitting	Fit	Fit
Checkpoint	Val	Same checkpoint
Pattern selection	–	Val subset
Group voting	–	Fit
Evaluation	Test	Test

`P` emits attributed paths: every simple path of one to L edges, written as its sequence of node types and canonicalized by taking the smaller of the sequence and its reverse. Path descriptors encode typed paths rather than rooted branching neighborhoods.

`C` emits ring patterns: for each cycle-basis ring of size three to eight, the pair of its size and the sorted multiset of its node types, together with the number of rings in each fused component. These ring descriptors complement the rooted unfoldings and paths represented by `W` and `P`.

The vocabulary is built on the mining set alone and then frozen. Other graphs use the frozen vocabulary, dropping unseen patterns without adding validation or test structures.

Matching relations. The BA2MOTIFS comparison varies ordinary and induced subgraph containment while keeping the supplied motifs, fitting indices, and base predictions fixed. Both relations respect node types; induced matching additionally requires the mapped vertices to have exactly the motif's adjacencies. The MUTAGENICITY supplied rule instead matches exact descriptor sets, and the mined application matches exact keys of selected-pattern indicators. These rules transfer labels only through their specified relation; their voting and fallback conditions are given below.

C.5 Pattern Selection, Voting, and Gating

Mined application. The prefilter settings θ_{cand} specify candidate support and purity thresholds, a ranking reference, and a pool size: the filter ranks patterns by single-pattern gain and keeps at most 300 candidates. The default candidate support is five selection graphs. Groups require at least three labeled examples, a unique majority, and smoothed purity $(\max_c n_{t,c} + 1)/(n_t + 2)$ at least as large as the chosen reference. These group checks apply when evaluating selection gain and when building fitting-label votes. Greedy selection stops at an improvement below 0.002, breaks ties by the smaller pattern identifier, and uses the group cap $M = \max(2, \lfloor |D_{\mathrm{vote}}|/2 \rfloor)$ through the implementation's conservative check $2g \leq M$, where g is the current number of selection groups. The selection objective $F(S)$ is evaluated on the selection set. The main application uses exact equality of the selected-pattern keys. If greedy selection returns no patterns, it retains the baseline everywhere. For a nonempty selected vocabulary, the key containing none of its patterns is a valid group and receives the same eligibility checks. The main multi-task application chooses (τ, K) per run by validation gain over $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.85, 0.9\}$ and $K \in \{1, 2, 3, 5, 10, 20\}$, breaking ties toward the smaller cap. It uses the GNN's validation accuracy as the reference at both selection and voting. The fixed choice $(0.50, 10)$ is a separate parameter-sensitivity comparison. The gate is a form of selective prediction (Geifman & El-Yaniv, 2017): a rejected group vote retains the trained GNN prediction. Here τ is the candidate-prefilter cut; selection and voting compare group purity with the GNN validation accuracy. The reference comparisons instead use a fixed threshold, or gate on a fixed threshold while scoring against the model. The pooled-ERM numerical study uses the exact fitting objective in Section 3.

BA2MOTIFS matching. At $t_n = 2$, the rule aggregates labels from all matching fitting explanations. A unique majority replaces the GNN prediction; no match or a tie retains it. Support and purity gates are disabled, so only the containment relation changes in Table 8. The supplied-motif learning curves require at least three voting graphs per group and a unique majority, without an additional purity threshold; Appendix D.1 distinguishes their coverage from that of the ungated matching test.

Supplied descriptors on MUTAGENICITY. The supplied rule groups fitting graphs by exact descriptor sets and counts their labels. For each test graph, a unique majority in the corresponding

fitting group replaces the GNN prediction. An absent group or a tied majority retains the GNN prediction. No extra support or purity threshold is imposed; a single fitting graph can supply a label.

Controlled explanation-class sweep. We evaluate GCN and GIN on three stratified splits and seeds 42–51 at per-class fitting budgets $t_n \in \{2, 5, 10, 20, 50, 100, 200, \text{Full}\}$. The rule uses each trained backbone as a fixed fallback predictor. The fitting subsets follow Appendix C.2, so each reported mean contains 30 runs. The structural application uses the ring-pattern miner, fitting labels for group votes, and the corresponding frozen GNN for fallback predictions. For this controlled sweep, validation gain selects $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.85, 0.9\}$ and the pattern cap $K_{\text{sel}} \in \{1, 2, 3, 5, 10, 20\}$, breaking a tie in favor of the smaller cap. The same grid is used in the main mined-pattern application.

C.6 Controls and Ablation Protocols

Main application controls. The permutation, random-pattern, and graph-size controls retain the mining set, selection objective, group eligibility checks, and fallback GNN wherever applicable, changing the representation or pattern choice specified below. The `perm` control permutes graph-to-pattern assignments to break their association with the labels while retaining the pattern vocabulary. The `randK` control draws K patterns uniformly from the vocabulary instead of selecting them for predictive gain. The `size` control forms keys from node count, edge count, or both. These controls test whether gains depend on structural associations, pattern selection, or graph size.

Random background subgraphs on MUTAGENICITY. The control denoted R in Table 10 replaces supplied patterns with random background subgraphs. For each supplied pattern eligible for subgraph matching, the control draws a connected subgraph outside the fitting graph's annotated explanation edges, with the requested edge count matched to that pattern and capped at eight. The large polycyclic descriptor is excluded from subgraph matching. Duplicate random patterns are combined and at most 300 distinct patterns are retained, ordered by their vote counts. The random patterns match a query by ordinary subgraph containment with node types respected; their fitting labels are pooled, with a tied or absent vote falling back to the same GNN. Thus, this condition changes both the structural representation and the matching rule relative to the supplied exact-set predictor. Its margin cannot isolate the effect of changing only the explanation source or only the matching relation. This control is evaluated only for fitting sets of at most 500 graphs, so no Full result is available.

Permutation with exact-set matching retained. Table 11 compares the supplied predictor with a permutation control that retains exact-set matching. For each seed and budget, the descriptor sets are randomly reassigned across the complete dataset, and the same exact-set majority rule is applied with the same fitting labels and frozen GNN. The permutation uses no held-out labels. It preserves the global multiset of descriptor sets and their frequencies, while breaking their association with the graphs. It need not preserve the realized fitting-group sizes or test coverage, because the permutation crosses split boundaries. The random generator is seeded by $7919s + t_n$ for finite budgets and $7919s + 999$ for Full, where s is the run seed.

Component comparisons. Component ablations use BA2MOTIFS, FOUR-MOTIF, AIDS, TOX21, DHFR, and PROTEINS with four backbones at $t_n = 50$, giving 24 dataset and backbone settings. Each comparison changes pattern choice, classifier form, or abstention while holding the other components and base predictions fixed. Pattern selection is compared with random patterns using the same unrestricted decision tree without a gate. The classifier comparison then contrasts a tree and group voting on the same selected-pattern indicators under the same gate. Five-fold logistic regression on node and edge counts, node-type histograms, and their combination checks the explanation-class family for simple label shortcuts.

Parameter and pattern-occurrence studies. The gate study compares the references in Appendix C.5, varies minimum support from 3 to 1, and substitutes miner `W` for miner `C`. Parameter selection compares per-seed validation selection over 36 (τ, K) configurations with fixed constants. In the leave-one-dataset-out check, the constant is selected using only validation gains from the other five datasets and then evaluated on the held-out dataset. A test-selected oracle is reported only as a retrospective comparison within this grid. The pattern-occurrence diagnostic uses BA2MOTIFS and FOUR-MOTIF with four backbones, eight budgets, three splits, and ten seeds, giving 1,920 runs. Validation gain selects the configuration from the same 6×6 grid of candidate purity thresholds and pattern caps used in the controlled explanation-class sweep, breaking ties toward the smaller cap. The voting reference is the GNN validation accuracy.

C.7 Metrics and Statistical Analysis

Accuracy differences and paired comparisons. All gains are accuracy differences in percentage points between paired predictions from the same run. We report hybrid-minus-GNN, control-minus-GNN, and hybrid-minus-control so that a harmful control cannot be mistaken for a useful hybrid. W/L/T counts positive, negative, and zero paired differences; tables headed W/L omit ties. The contrast is identified in each caption, since superiority to a control does not establish improvement over the GNN. The control margin compares the rule with the better control within each run:

$$\Delta_{\mathrm{ctrl}} = \mathrm{mean}_r \left[\mathrm{Acc}_r(h_{\mathrm{rule}}) - \max\{\mathrm{Acc}_r(h_{\mathrm{perm}}), \mathrm{Acc}_r(h_{\mathrm{randK}})\} \right].$$

Size is compared in a separate column. The mean of this paired margin can differ from the difference between mean accuracies because the better control changes with the seed and fitting subset.

Aggregation and significance. Three-split curves average ten seeds per split; their bands span the three split means and are not confidence intervals. Original-split comparisons average ten paired seeds. For these comparisons, the two-sided sign-flip test (Winkler et al., 2014) enumerates all $2^{10} = 1{,}024$ sign assignments and compares the absolute permuted sum with the observed sum; there is no adjustment across budgets. The strict-budget and kernel comparisons use the procedures and multiplicity corrections specified in their table captions. The explanation-class sweep has no separate significance test. Sign-flip inference assumes that the joint null distribution of paired differences is invariant under the permitted flips. The graph curves do not report confidence intervals or standard errors; uncertainty estimates for the numerical studies are described separately in Appendix B.

Explanation-class gains and headroom. Let $A_{\mathrm{GNN}}(L, t_n)$ and $A_{\mathrm{hyb}}(L, t_n)$ be the mean test accuracies in percentage points for the paired GNN and hybrid runs, and define

$$\Delta(L, t_n) = A_{\mathrm{hyb}}(L, t_n) - A_{\mathrm{GNN}}(L, t_n). \tag{28}$$

For a target gain $g \in \{2, 5, 10\}$, Table 9 reports the smallest measured t_n for which $\Delta(L, t_n) \geq g$; a dash means that none of the eight budgets reaches the target. We normalize the gain by the GNN error to measure captured accuracy headroom:

$$H(L, t_n) = \frac{\Delta(L, t_n)}{100 - A_{\mathrm{GNN}}(L, t_n)}. \tag{29}$$

For example, $H = 1$ means that the hybrid closes the entire gap between the GNN and perfect accuracy, while $H = 0$ means no improvement.

Coverage and group diagnostics. Coverage is the fraction of test graphs receiving a rule prediction instead of GNN fallback. We count formed groups as distinct vote-set keys, eligible groups as those passing support, unique-majority, and purity checks, and fired groups as those predicting at least one test label. Uncovered queries receive the first applicable rejection reason: absent key, insufficient support, tied vote, or insufficient purity. Coverage and these rejection fractions sum to one.

Agreement of pattern and motif occurrences. For a pattern p and a planted motif m, let A_p and A_m be the sets of graph indices carrying them. Their occurrence agreement is $|A_p \cap A_m|/|A_p \cup A_m|$. We report the maximum over all candidate-pattern–motif pairs and the corresponding maximum restricted to selected patterns, assigning zero to an empty selection. These scores measure agreement of graph occurrences, not subgraph shape or recovery of every motif. They are evaluated over the full dataset after selection, including the test graphs, solely for diagnosis; neither selection nor prediction uses these scores.

C.8 Comparison Protocol for GSAT and DIR

The comparison in Appendix D.8 evaluates GNN, Hybrid, GSAT, and DIR on BA2MOTIFS, FOUR-MOTIF, AIDS, TOX21, DHFR, and PROTEINS with GCN and GIN. It uses the three splits, ten seeds per split, and eight fitting budgets in Appendix C.2, giving 2,880 matched evaluations across tasks, backbones, and budgets. Within each setting, all methods use the same fitting graphs, validation set, and test set. The fitting budget excludes the shared validation labels: all methods use validation for checkpoint selection, and Hybrid also uses a subset of these labels for pattern selection (Table 7).

Training and explanation use. The GNN is trained by cross-entropy minimization, and Hybrid applies miner C and group voting to its frozen predictions with $(\tau, K) = (0.50, 10)$ and the model-referenced gate. GSAT learns stochastic edge attention through an information-bottleneck objective

(Miao et al., 2022), while DIR learns an explanatory subgraph through intervention-based training (Wu et al., 2022). Both are reimplemented with the encoder depth, hidden width, readout, optimizer, and training duration of Appendix C.3. Their method-specific parameters are fixed across tasks without a hyperparameter search.

GSAT uses a Bernoulli reference rate of 0.7, information-loss weight 1.0, and sampling temperature 1.0; the reference rate is constant and there is no pretraining phase. DIR reduces the retained-edge ratio from 1.0 to 0.5 over the first half of training and uses four within-batch interventions with weight 0.5 on the sum of their mean loss and loss variance. These interventions combine the explanatory representation with a shuffled complementary representation, with gradients stopped through the complementary branch. DIR evaluates the selected checkpoint at retained-edge ratio 0.5. These settings specify the evaluated implementations; their results do not establish performance under separately optimized configurations.

Reporting. All four methods have results for every matched setting. Each task–backbone–budget accuracy averages thirty runs, and table subscripts give the sample standard deviation across those runs. Curve bands span the three split means and are not confidence intervals. Aggregate gains average the twelve task–backbone means with equal weight. All budgets are reported, including $t_n = 2$ and Full; no budget is selected using test accuracy.

D Additional Graph Results

This appendix reports the matching, supplied-descriptor, and mined-rule graph experiments. Their protocols are given in Section 4 and Appendix C.

D.1 Matching and Explanation-Group Complexity

Table 8 holds the supplied motifs and trained GNN fixed while changing only containment. A five-cycle is an ordinary subgraph of the house, whose induced graph on those vertices has six edges. Induced matching distinguishes these two supplied keys on this support. Support and purity gates are disabled in this test. The separate supplied-motif learning curves require at least three voting examples per group and thus have different small-budget coverage.

Table 8: BA2MOTIFS at two fitting graphs per class. Entries are accuracy fractions, with standard deviation over ten seeds on one split. OM and IM denote ordinary and induced matching.

Model	GNN	OM	IM
GIN	$0.636_{\pm 0.155}$	$0.792_{\pm 0.113}$	$1.000_{\pm 0.000}$
GCN	$0.500_{\pm 0.015}$	$0.671_{\pm 0.111}$	$1.000_{\pm 0.000}$

Group-count sweep. All possible motif keys occur in the generated family. Five-fold logistic regression obtains AUC 0.489–0.510 from node and edge counts, 0.487–0.508 from node-type histograms, and 0.484–0.508 from both together. These checks support the intended control of simple size and type-count shortcuts. At $t_n = 20$, GCN coverage over the five group counts is $0.974, 0.726, 0.680, 0.549, 0.496$, while GIN coverage is $0.974, 0.704, 0.638, 0.533, 0.482$. The corresponding gains relative to the GNN's remaining error appear in Table 9, which also gives the smallest fitting budget at which the hybrid reaches each gain; the budget for a ten-point gain grows from five to one hundred examples per class as J grows from four to thirty-seven. This relationship between group complexity, coverage, and gain provides a qualitative check of the discovery mechanism. The selected keys may differ from the full planted-key map, so this is a qualitative class-count comparison.

D.2 Supplied Descriptors on MUTAGENICITY

Ten predefined descriptors approximate toxicophore structures using atom types and adjacency. Graphs with identical descriptor sets share fitting labels, with a unique majority replacing the GNN and absent or tied groups retaining it. There is no additional support or purity threshold. Table 10

Table 9: Smallest per-class fitting budget at which the hybrid reaches the indicated accuracy gain over the GNN on the explanation-class family without label noise. HR_{20} is the fraction of GNN error removed at $t_n = 20$. A dash means the target was not reached on the measured grid.

		GCN gain			GIN gain			HR_{20}	
L	J	$+2$	$+5$	$+10$	$+2$	$+5$	$+10$	GCN	GIN
2	4	5	5	5	5	5	5	0.956	0.959
4	11	10	10	20	10	10	20	0.439	0.416
6	22	10	10	50	10	20	50	0.159	0.142
8	37	20	50	100	20	50	100	0.087	0.079
10	56	50	100	–	20	–	–	0.043	0.050

gives every fitting budget. The supplied predictor improves over the GNN in thirteen of fourteen finite-budget settings, with gains of 5.15 points for GIN and 7.23 for GCN at five fitting graphs per class. The advantage in these results occurs at finite fitting budgets.

The random-subgraph comparison changes both representation and matching. Table 11 instead retains exact-set matching and permutes descriptor assignments. The supplied rule exceeds that permutation control in all sixteen model–budget combinations. The reported sign-flip tests compare supplied versus permuted assignments, not improvement over the GNN. A control's loss can contribute to the source–control margin, so the gain over the GNN remains the primary quantity.

Table 10: MUTAGENICITY: supplied descriptor sets (S) against the same trained GNN (G) and a random background-subgraph control (R). Accuracy points over ten paired seeds on the original split, S−G subscripted with its standard deviation over those seeds; W/L and the exact sign-flip p compare S with R, not S with G. R was not recorded at the full pool.

	S−G		R−G		S−R		W/L		p	
t_n	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN
2	$+4.23_{\pm 4.71}$	$+5.22_{\pm 4.99}$	-5.52	-5.82	+9.75	+11.04	9/1	9/1	0.004	0.004
5	$+5.15_{\pm 3.25}$	$+7.23_{\pm 5.01}$	-9.24	-8.34	+14.39	+15.57	10/0	10/0	0.002	0.002
10	$+2.54_{\pm 4.66}$	$+2.77_{\pm 3.69}$	-9.86	-10.16	+12.40	+12.93	10/0	10/0	0.002	0.002
20	$+2.08_{\pm 3.95}$	$+3.60_{\pm 3.35}$	-12.86	-10.55	+14.94	+14.16	10/0	10/0	0.002	0.002
50	$+1.73_{\pm 2.55}$	$+3.05_{\pm 3.60}$	-14.18	-13.16	+15.91	+16.21	10/0	10/0	0.002	0.002
100	$+2.12_{\pm 3.40}$	$+1.06_{\pm 2.61}$	-15.33	-16.26	+17.46	+17.32	10/0	10/0	0.002	0.002
200	$-0.07_{\pm 1.47}$	$+0.35_{\pm 0.78}$	-18.34	-18.18	+18.27	+18.52	10/0	10/0	0.002	0.002

Table 11: MUTAGENICITY supplied descriptor sets (S) versus permuted descriptor-set assignments (P), with exact-set matching and the same trained GNN (G) retained. Each row is one per-class fitting budget, and every quantity is reported for GIN and for GCN. Differences are accuracy points averaged over ten paired seeds on the original split, S−G subscripted with its standard deviation over those seeds. W/L and p test S versus P, not S versus G.

	S−G		P−G		S−P		W/L		p	
t_n	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN
2	$+4.23_{\pm 4.71}$	$+5.22_{\pm 4.99}$	-4.64	-4.62	+8.87	+9.84	10/0	10/0	0.002	0.002
5	$+5.15_{\pm 3.25}$	$+7.23_{\pm 5.01}$	-6.54	-5.47	+11.69	+12.70	10/0	10/0	0.002	0.002
10	$+2.54_{\pm 4.66}$	$+2.77_{\pm 3.69}$	-9.56	-9.45	+12.10	+12.22	10/0	10/0	0.002	0.002
20	$+2.08_{\pm 3.95}$	$+3.60_{\pm 3.35}$	-15.27	-13.37	+17.34	+16.97	10/0	10/0	0.002	0.002
50	$+1.73_{\pm 2.55}$	$+3.05_{\pm 3.60}$	-16.61	-15.68	+18.34	+18.73	10/0	10/0	0.002	0.002
100	$+2.12_{\pm 3.40}$	$+1.06_{\pm 2.61}$	-18.01	-18.59	+20.14	+19.65	10/0	10/0	0.002	0.002
200	$-0.07_{\pm 1.47}$	$+0.35_{\pm 0.78}$	-21.11	-20.83	+21.04	+21.18	10/0	10/0	0.002	0.002
Full	$-4.18_{\pm 1.10}$	$-4.00_{\pm 0.97}$	-23.74	-23.51	+19.56	+19.52	10/0	10/0	0.002	0.002

D.3 Complete Real-Task Curves

Figure 8 shows the 128 real-task settings: four tasks, four backbones, and eight fitting budgets. Each point averages ten seeds on each of three splits. It reports mean gains, coverage, pattern controls, and the graph-size control. The decomposition in Table 12 and the component analyses use the same sweep. Their gains and coverage agree with the curves to the reported precision.

Table 12 fixes $t_n = 50$ rather than selecting each task's best test budget. Pattern controls and the graph-size control are retained in that table. The size control is competitive on AIDS, so the attribution of its improvement to structural descriptors remains limited. Larger fitting budgets change the base predictor as well as the available group labels; they need not produce a monotone hybrid gain.

Table 12: Real tasks at $t_n = 50$. Rule is the group vote alone, with the training majority where it abstains; Hybrid keeps the GNN there instead. Gain is Hybrid minus GNN, subscripted with the standard deviation of the three split means; Cov the fraction of test graphs decided, Δ_{ctrl} the margin over the better pattern control, Size the graph-size control's gain, and W/L wins and losses over 30 runs. SAGE denotes GraphSAGE.

	AIDS								DHFR							
Model	GNN	Rule	Hybrid	Gain	Cov	Δ_{ctrl}	Size	W/L	GNN	Rule	Hybrid	Gain	Cov	Δ_{ctrl}	Size	W/L
GIN	86.93	74.17	91.05	$+4.12_{\pm 0.83}$	0.51	+4.05	+5.00	25/0	69.78	57.42	71.38	$+1.60_{\pm 1.73}$	0.23	+1.29	-0.13	14/5
GCN	73.07	76.73	81.15	$+8.08_{\pm 1.88}$	0.73	+7.20	+9.22	26/3	68.89	57.27	70.00	$+1.11_{\pm 2.02}$	0.26	+0.62	+0.04	14/7
PNA	78.83	78.75	84.77	$+5.93_{\pm 0.70}$	0.65	+5.53	+8.70	27/1	73.20	54.53	72.13	$-1.07_{\pm 1.78}$	0.15	-1.20	+0.00	5/5
SAGE	73.55	77.12	82.08	$+8.53_{\pm 1.83}$	0.75	+7.87	+7.30	30/0	69.02	56.93	69.38	$+0.36_{\pm 2.33}$	0.26	+0.09	-0.09	13/8
	TOX21								PROTEINS							
Model	GNN	Rule	Hybrid	Gain	Cov	Δ_{ctrl}	Size	W/L	GNN	Rule	Hybrid	Gain	Cov	Δ_{ctrl}	Size	W/L
GIN	64.39	58.31	67.70	$+3.32_{\pm 0.60}$	0.26	+2.95	-0.02	27/0	65.86	54.61	66.40	$+0.54_{\pm 1.32}$	0.21	+0.48	+0.00	14/6
GCN	61.70	59.13	66.37	$+4.67_{\pm 1.34}$	0.34	+3.88	-0.09	28/2	64.43	56.47	66.22	$+1.79_{\pm 1.21}$	0.27	+1.76	+0.00	18/5
PNA	66.25	57.86	68.44	$+2.19_{\pm 0.39}$	0.24	+1.79	-0.09	23/2	64.55	55.12	65.57	$+1.01_{\pm 1.47}$	0.25	+0.86	+0.00	16/9
SAGE	63.00	59.13	67.33	$+4.33_{\pm 1.60}$	0.35	+3.65	-0.23	29/1	64.02	55.67	64.94	$+0.92_{\pm 0.62}$	0.25	+0.83	+0.00	19/5

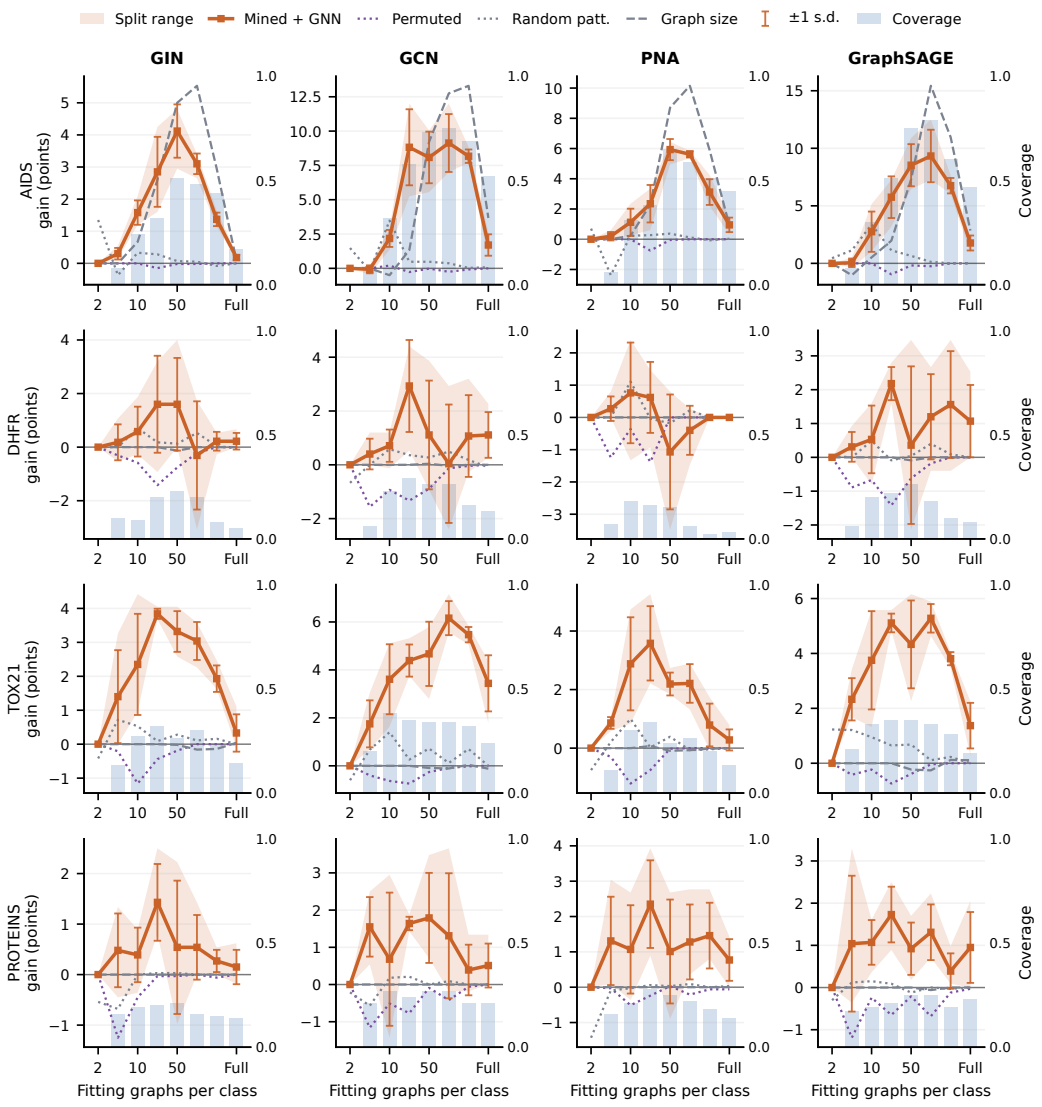


Figure 8: Gain over the same trained GNN on the four real tasks (rows), four backbones (columns), and eight fitting budgets ending at the full pool. Orange curves show the mined-rule hybrid and the other curves are controls; light blue bars give coverage on the right axes. Each point averages thirty runs, with error bars for the standard deviation of the three split means and shaded bands for their range. Gain-axis limits differ between panels.

D.4 Strict Total Label Budgets

Every training, checkpoint, selection, and voting label counts against a single budget $B \in \{100, 200, 500, 1000\}$. Labels are drawn from the merged original fitting and validation pool in its class proportions. A stratified $v = \min(50, \lceil 0.2B \rceil)$-label subset serves checkpoint selection; all other labels train the GNN. Pattern mining uses training graphs, selection reuses checkpoint labels, and voting reuses training labels. The same fitted GNN is the hybrid's fallback. DHFR has only 681 labels in this pool and uses all of them at the largest budget. Each cell contains three splits and ten seeds.

Table 13 gives all results. Thirty-one of the thirty-two mean accuracy gains are positive, and seven of the eight TOX21 accuracy gains are significant.

Table 13: Complete strict-budget results, in percentage points. Cov is the fraction of test graphs receiving a structural vote, BA denotes balanced accuracy, and W/L counts paired wins and losses over 30 runs, omitting ties. An asterisk marks a gain significant at the 5% level in a paired sign-flip test, Holm-corrected within each backbone's sixteen cells for each metric. Gains use unrounded run differences and may differ by 0.01 from subtraction of the displayed accuracies.

		GNN		Hybrid		Cov		ΔAcc (W/L)		ΔBA (W/L)	
	B	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN
AIDS	100	77.82	83.13	78.77	83.78	.46	.27	$+0.95^*$ (9/0)	$+0.65$ (7/2)	$+0.38$ (8/2)	-0.94 (6/3)
	200	77.88	87.70	80.72	89.63	.54	.35	$+2.83^*$ (22/0)	$+1.93^*$ (21/0)	$+1.65^*$ (21/1)	$+1.08^*$ (20/1)
	500	81.12	92.27	84.57	93.05	.50	.23	$+3.45^*$ (29/0)	$+0.78^*$ (15/0)	$+2.28^*$ (28/1)	$+0.49^*$ (15/0)
	1000	85.25	94.72	87.52	94.92	.49	.07	$+2.27^*$ (25/0)	$+0.20$ (3/0)	$+0.76$ (24/3)	$+0.12$ (3/0)
DHFR	100	63.73	64.22	63.60	65.20	.14	.09	-0.13 (4/3)	$+0.98$ (6/1)	-0.68 (4/4)	$+0.65$ (6/1)
	200	69.42	69.33	70.71	71.07	.18	.14	$+1.29^*$ (15/3)	$+1.73^*$ (15/1)	$+0.90$ (14/7)	$+1.18^*$ (14/4)
	500	74.27	77.78	75.56	78.27	.19	.14	$+1.29^*$ (15/5)	$+0.49$ (9/5)	$+0.67$ (15/7)	$+0.25$ (9/5)
	1000	75.69	80.76	76.31	80.84	.12	.05	$+0.62$ (10/5)	$+0.09$ (2/2)	$+0.27$ (10/6)	-0.03 (2/3)
TOX21	100	59.40	59.32	62.40	61.74	.23	.15	$+3.00^*$ (19/0)	$+2.42^*$ (14/1)	$+3.00^*$ (19/0)	$+2.42^*$ (14/1)
	200	61.75	62.84	65.49	65.00	.20	.12	$+3.74^*$ (21/0)	$+2.16^*$ (16/1)	$+3.74^*$ (21/0)	$+2.16^*$ (16/1)
	500	65.61	69.84	68.65	71.09	.21	.13	$+3.04^*$ (23/0)	$+1.25^*$ (16/1)	$+3.04^*$ (23/0)	$+1.25^*$ (16/1)
	1000	67.49	73.18	70.46	73.42	.14	.08	$+2.96^*$ (21/0)	$+0.25$ (8/2)	$+2.96^*$ (21/0)	$+0.25$ (8/2)
PROTEINS	100	63.63	63.48	64.85	63.93	.15	.07	$+1.22$ (7/2)	$+0.45$ (2/2)	$+1.20$ (7/2)	$+0.49$ (2/4)
	200	64.67	66.07	65.74	66.79	.21	.17	$+1.07$ (10/4)	$+0.71$ (6/4)	$+1.16$ (11/8)	$+0.71$ (6/9)
	500	66.34	68.30	67.71	68.81	.26	.12	$+1.37^*$ (15/3)	$+0.51$ (10/2)	$+1.25$ (16/4)	$+0.29$ (10/2)
	1000	67.23	70.42	67.65	70.51	.20	.06	$+0.42$ (9/4)	$+0.09$ (4/2)	$+0.34$ (10/8)	$+0.00$ (5/3)

D.5 Sharing Labels with a Kernel Predictor

The kernel comparison uses a Weisfeiler–Leman subtree kernel with a precomputed-kernel SVM. Its depth $h \in \{1, 2, 3\}$ and regularization $C \in \{0.01, 0.1, 1, 10, 100\}$ are selected by validation accuracy. Structural selection is rerun with this predictor's validation accuracy as reference. The left block of Table 14 uses the original fitting and validation protocol, with no GNN or kernel-confidence gate. Its rule coverages on AIDS, DHFR, TOX21, and PROTEINS are $0.59, 0.20, 0.18, 0.18$.

The right block splits the fitting subset by class, using one half for kernel training and pattern mining and the other for votes. Validation is also split: one half selects kernel parameters and supplies the reference accuracy, and the other selects patterns. Overrides are restricted to keys realized on the pattern-selection half. The split seeds are the run seed plus 4242 for fitting and plus 5353 for validation. These are comparisons with each protocol's own fitted kernel, not between kernels trained on unequal numbers of labels.

On TOX21, graphs carrying none of the selected ring patterns form a key covering about 18% of fitting graphs. It votes inactive in all 30 runs and passes the gate in 27. Its shared-vote error is 0.116 and the kernel error is 0.181 on the 959 covered test-graph instances pooled across those runs. These are repeated test instances across runs, not 959 independent graphs. The diagnostic identifies the group carrying the observed improvement.

Table 14: Structural label sharing with a WL-kernel SVM at $t_n = 50$, without a GNN. The right block splits fitting data between kernel training and rule voting. Accuracies are in percent and gains in points with paired wins and losses; an asterisk marks a gain significant at the 5% level in a paired sign-flip test, Holm-corrected across the four tasks.

	Shared fitting sample			Split fitting sample		
Task	WL	WL+rule	Gain (W/L)	WL	WL+rule	Gain (W/L)
AIDS	85.42	90.35	$+4.93^*$ (29/0)	79.58	86.35	$+6.77^*$ (29/1)
DHFR	75.47	74.49	-0.98 (9/6)	67.60	68.58	$+0.98$ (14/4)
TOX21	68.91	70.14	$+1.23^*$ (24/3)	64.95	66.81	$+1.86^*$ (24/3)
PROTEINS	67.50	68.15	$+0.65$ (12/4)	66.67	68.04	$+1.37$ (11/3)

D.6 A TUNED BASE NETWORK

The base GNNs of the main sweep share one fixed training recipe. This study replaces it with a validation search at $t_n = 50$ on the grid of Table 12, with four tasks, four backbones, and thirty runs per cell. In every run each backbone is trained under twenty-five configurations, the recipe and twenty-four random draws over learning rate, batch size, epochs, width, depth, dropout, weight decay, and mean or sum readout, and the configuration with the highest validation accuracy is kept; test labels take no part in the choice. The rule, its selection, and the statistics are unchanged, the gate reference becomes the tuned network's validation accuracy, and the recipe network is retrained alongside, so every comparison is paired within one training environment. Conditions and decision rules were fixed before any training.

Table 15 reports the outcome. Tuning raises the network in all sixteen cells. On TOX21 it adds 0.5 to 3.3 points to the network, and the hybrid still gains 4.61, 3.11, 4.18, and 1.32 points over the tuned network for GCN, GIN, GraphSAGE, and PNA, each significant after Holm correction and each above both the graph-size and the pattern controls. On AIDS the search selects sum readout in all 120 runs and the tuned networks reach 96 to 99%, leaving the gate almost nothing to correct, with coverage at most 0.08. On DHFR and PROTEINS the hybrid stays within half a point of the tuned networks, and no cell is significantly negative. The reference is the highest of twenty-five validation accuracies, which exceeds test accuracy by 7 to 12 points on TOX21, DHFR, and PROTEINS and makes the gate stricter.

Table 15: A tuned base network at $t_n = 50$. GNN is the paper's recipe and Tuned the best of twenty-five configurations on validation accuracy, both retrained for this study, so GNN and the recipe gain Hyb.−GNN differ slightly from Table 12. Hyb.$_{\mathrm{T}}$ runs the rule on the tuned network with its validation accuracy as the reference. Differences are means in points over thirty paired runs, with wins and losses; an asterisk marks significance at the 5% level in a paired sign-flip test, Holm-corrected over the sixteen cells. Cov gives the coverage on the recipe and on the tuned network, and SAGE denotes GraphSAGE.

	AIDS						DHFR					
Model	GNN	Tuned	Hyb.$_{\mathrm{T}}$	Hyb.$_{\mathrm{T}}$−Tuned	Hyb.−GNN	Cov	GNN	Tuned	Hyb.$_{\mathrm{T}}$	Hyb.$_{\mathrm{T}}$−Tuned	Hyb.−GNN	Cov
GIN	85.45	99.22	99.25	+0.03 (1/0)	+4.55* (26/0)	0.52/0.02	69.02	72.49	72.98	+0.49 (7/1)	+1.56 (13/5)	0.20/0.07
GCN	73.07	96.43	96.73	+0.30 (4/0)	+8.17* (27/3)	0.73/0.08	67.91	69.42	69.11	−0.31 (9/6)	+1.73 (17/6)	0.28/0.20
PNA	78.17	97.98	97.98	+0.00 (0/0)	+6.27* (27/1)	0.64/0.00	73.16	74.84	75.24	+0.40 (3/0)	−0.80 (6/5)	0.14/0.02
SAGE	73.57	98.03	98.22	+0.18 (1/0)	+8.53* (30/0)	0.74/0.03	69.29	71.91	72.09	+0.18 (8/2)	+0.22 (12/8)	0.25/0.10
	TOX21						PROTEINS					
Model	GNN	Tuned	Hyb.$_{\mathrm{T}}$	Hyb.$_{\mathrm{T}}$−Tuned	Hyb.−GNN	Cov	GNN	Tuned	Hyb.$_{\mathrm{T}}$	Hyb.$_{\mathrm{T}}$−Tuned	Hyb.−GNN	Cov
GIN	62.86	64.91	68.02	+3.11* (27/0)	+4.14* (27/1)	0.27/0.20	65.65	68.10	67.92	−0.18 (5/5)	+0.62 (13/6)	0.21/0.11
GCN	60.93	61.46	66.07	+4.61* (28/1)	+4.46* (27/2)	0.32/0.27	64.76	68.18	68.12	−0.06 (3/3)	+1.34* (20/5)	0.27/0.07
PNA	66.25	69.54	70.86	+1.32* (21/3)	+2.65* (26/0)	0.26/0.17	64.49	68.07	68.24	+0.18 (6/3)	+1.61* (19/4)	0.26/0.05
SAGE	62.81	63.91	68.09	+4.18* (29/0)	+4.65* (30/0)	0.31/0.23	64.85	67.38	67.44	+0.06 (5/4)	+1.10 (16/3)	0.27/0.09

A second study tunes the network under the strict budget at $B = 500$ and 1000, over four configurations, the recipe and the three configurations most often selected in the earlier tuning studies. Validation is nested: half of the checkpoint labels choose the configuration and checkpoint, and the other half set the gate reference. Table 16 reports the outcome. The deployed hybrid keeps significant gains over the tuned network in all four TOX21 settings, 0.93 to 3.18 points. On AIDS the tuned network already reaches 94 to 97%, leaving at most 0.65 points to gain.

Table 16: A tuned network under the strict budget. Recipe and Tuned are test accuracies in percent of the recipe network and the network tuned over four configurations with nested validation. Hyb.−Tuned gives the deployed hybrid's gain over the tuned network with wins and losses, Hyb.−Recipe its gain over the recipe network, and BA the balanced-accuracy version of Hyb.−Tuned, all in points.

		Recipe		Tuned		Tuned−Recipe		Hyb.−Tuned (W/L)		Hyb.−Recipe		BA	
	B	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN
TOX21	500	63.88	69.26	64.82	68.58	+0.95	−0.68	+3.18* (23/4)	+1.56* (21/4)	+4.44*	+1.60*	+3.18*	+1.56*
	1000	67.75	72.11	68.19	71.51	+0.44	−0.60	+2.35* (24/0)	+0.93* (15/3)	+3.75*	+1.02*	+2.35*	+0.93*
AIDS	500	80.17	88.28	96.82	94.40	+16.65*	+6.12*	+0.10 (2/0)	+0.65 (5/1)	+3.35*	+2.25*	+0.06	+0.12
	1000	83.85	91.53	95.63	94.38	+11.78*	+2.85*	+0.40 (3/1)	+0.60* (12/1)	+2.27*	+1.20*	+0.22	−0.16
DHFR	500	72.71	77.56	72.22	76.80	−0.49	−0.76	+0.22 (15/10)	−1.20 (8/8)	−0.13	−1.20	−1.58	−1.89
	1000	73.91	80.09	75.16	80.09	+1.24	−0.00	−0.13 (9/9)	−1.33 (8/8)	−0.49	−1.33	−0.21	−1.79
PROTEINS	500	65.36	67.95	68.21	68.84	+2.86*	+0.89	+1.10 (15/8)	+0.71 (18/4)	+1.73	+0.62	+0.86	+0.28
	1000	66.19	70.00	69.11	70.80	+2.92*	+0.80	+0.74 (12/7)	−0.27 (9/7)	+0.83	−0.42	+0.09	−0.94

D.7 Component and Parameter Comparisons

Table 17 gives component comparisons at $t_n = 50$ across six tasks and four backbones. The selected-versus-random comparison uses an unrestricted tree without a gate in both conditions. The tree-versus-vote comparison uses the same selected patterns and the same gate. The two gate comparisons replace a rejected structural prediction with the trained GNN. Different rows use different comparisons and their gains are not additive.

Table 17: Component ablations at $t_n = 50$, averaged over 24 dataset–backbone settings. Gains (percentage points) and W/L/T compare the first option with the second.

Comparison	Gain	W/L/T
Selected vs. random patterns	+16.42	22/2/0
Tree vs. group voting	+0.20	13/5/6
Group voting: gate on vs. off	+11.52	22/0/2
Tree: gate on vs. off	+7.89	17/5/2

The selection comparison wins twenty-two of twenty-four settings; the group-vote gate improves twenty-two and ties two. The gated tree and gated vote differ by only 0.20 points on average, with thirteen wins, five losses, and six ties for the tree. This supports the importance of structural selection and fallback rather than a general superiority of majority voting.

The leave-one-dataset-out parameter check selects a fixed configuration using validation gains on the other five datasets and evaluates it on the remaining dataset. All six folds select $(\tau, K) = (0.50, 10)$; per-seed validation selection exceeds this constant by 0.02 points on average. Table 18 reports all folds. This is evidence within the evaluated tasks and grid, not an assertion that the setting is optimal on other tasks.

Table 18: Leave-one-dataset-out choice of fixed parameters from validation gains. Accuracy is in percent; Difference is per-seed minus constant.

Dataset	Constant	Per-seed	Constant accuracy	Difference
AIDS	$\tau = 0.50$, $K = 10$	80.12	80.09	+0.03
DHFR	$\tau = 0.50$, $K = 10$	68.41	68.41	+0.00
TOX21	$\tau = 0.50$, $K = 10$	65.47	65.45	+0.01
PROTEINS	$\tau = 0.50$, $K = 10$	65.01	65.02	-0.02
BA2MOTIFS	$\tau = 0.50$, $K = 10$	94.39	94.39	+0.00
FOUR-MOTIF	$\tau = 0.50$, $K = 10$	80.87	80.80	+0.07

D.8 Comparison with Explanation-Learning Methods

GSAT (Miao et al., 2022) and DIR (Wu et al., 2022) learn explanatory representations together with a predictor. Here they are compared with the mined hybrid using the matched protocol in Appendix C.8. Results for every task–backbone–budget configuration appear in Tables 19 and 20. The evaluated reimplementations use fixed method-specific settings; the comparison does not claim separately optimized performance for each alternative.

At fifty fitting graphs per class, the hybrid's mean gain over the matched GNN is 9.64 points across the six tasks and two backbones. This aggregate includes the larger controlled-task improvements; per-task comparisons are the appropriate evidence for real-task performance. These methods share this comparison's fitting samples and validation protocol; its settings differ from the main sweep.

Table 19: Test accuracy (%) on the two controlled tasks for GNN, GSAT, DIR, and Hybrid under Appendix C.8. Entries are means over thirty runs (three splits, ten seeds each), with sample standard deviations as subscripts. Full uses the entire fitting pool.

		GIN				GCN			
	t_n	GNN	GSAT	DIR	Hybrid	GNN	GSAT	DIR	Hybrid
BA2MOTIFS	2	$68.60_{\pm 11.89}$	$67.90_{\pm 14.54}$	$51.60_{\pm 6.72}$	$68.60_{\pm 11.89}$	$50.13_{\pm 5.30}$	$50.23_{\pm 5.10}$	$54.77_{\pm 9.78}$	$50.13_{\pm 5.30}$
	5	$82.37_{\pm 10.58}$	$77.27_{\pm 14.80}$	$54.93_{\pm 15.05}$	$95.63_{\pm 6.04}$	$52.07_{\pm 5.60}$	$51.10_{\pm 5.27}$	$55.23_{\pm 10.99}$	$100.00_{\pm 0.00}$
	10	$94.87_{\pm 4.78}$	$92.57_{\pm 6.57}$	$60.27_{\pm 19.13}$	$97.10_{\pm 2.63}$	$51.90_{\pm 5.22}$	$50.93_{\pm 4.47}$	$57.63_{\pm 13.91}$	$100.00_{\pm 0.00}$
	20	$97.60_{\pm 3.11}$	$97.07_{\pm 4.78}$	$58.73_{\pm 17.94}$	$98.60_{\pm 1.07}$	$51.40_{\pm 5.68}$	$51.27_{\pm 4.79}$	$58.17_{\pm 14.94}$	$100.00_{\pm 0.00}$
	50	$98.63_{\pm 1.85}$	$98.73_{\pm 2.05}$	$56.23_{\pm 16.23}$	$99.10_{\pm 1.49}$	$51.23_{\pm 4.85}$	$51.40_{\pm 4.85}$	$56.20_{\pm 8.83}$	$100.00_{\pm 0.00}$
	100	$98.70_{\pm 1.53}$	$98.50_{\pm 2.10}$	$54.07_{\pm 12.92}$	$99.17_{\pm 0.87}$	$52.27_{\pm 4.18}$	$52.27_{\pm 3.96}$	$58.53_{\pm 10.95}$	$100.00_{\pm 0.00}$
	200	$99.67_{\pm 0.55}$	$99.77_{\pm 0.77}$	$51.13_{\pm 6.21}$	$99.67_{\pm 0.55}$	$51.80_{\pm 3.96}$	$50.97_{\pm 3.34}$	$53.87_{\pm 7.59}$	$100.00_{\pm 0.00}$
	Full	$99.50_{\pm 1.17}$	$99.20_{\pm 1.69}$	$51.53_{\pm 8.21}$	$99.50_{\pm 1.17}$	$50.17_{\pm 1.37}$	$50.37_{\pm 1.59}$	$49.60_{\pm 2.30}$	$100.00_{\pm 0.00}$
FOUR-MOTIF	2	$56.70_{\pm 4.71}$	$57.42_{\pm 5.75}$	$53.64_{\pm 6.81}$	$56.70_{\pm 4.71}$	$51.94_{\pm 2.46}$	$51.84_{\pm 2.97}$	$54.19_{\pm 3.39}$	$51.94_{\pm 2.46}$
	5	$58.67_{\pm 4.76}$	$60.12_{\pm 6.02}$	$52.23_{\pm 5.31}$	$65.16_{\pm 7.11}$	$52.75_{\pm 3.09}$	$54.60_{\pm 3.36}$	$55.70_{\pm 5.37}$	$62.84_{\pm 8.59}$
	10	$61.57_{\pm 5.07}$	$60.35_{\pm 5.29}$	$52.48_{\pm 5.69}$	$73.23_{\pm 8.04}$	$56.71_{\pm 4.10}$	$60.21_{\pm 5.41}$	$58.48_{\pm 4.72}$	$70.49_{\pm 9.37}$
	20	$63.84_{\pm 2.87}$	$65.05_{\pm 2.52}$	$57.12_{\pm 12.41}$	$76.91_{\pm 4.32}$	$64.93_{\pm 4.71}$	$67.57_{\pm 4.25}$	$64.53_{\pm 6.56}$	$77.81_{\pm 3.49}$
	50	$70.34_{\pm 4.01}$	$67.01_{\pm 3.82}$	$63.30_{\pm 16.47}$	$95.77_{\pm 4.18}$	$76.38_{\pm 4.06}$	$76.45_{\pm 4.78}$	$72.22_{\pm 7.44}$	$92.53_{\pm 6.99}$
	100	$81.31_{\pm 6.87}$	$73.86_{\pm 9.06}$	$58.45_{\pm 15.87}$	$91.67_{\pm 7.56}$	$85.41_{\pm 4.55}$	$83.15_{\pm 5.97}$	$78.33_{\pm 11.30}$	$87.36_{\pm 5.44}$
	200	$94.22_{\pm 2.36}$	$90.42_{\pm 4.71}$	$62.73_{\pm 19.45}$	$94.22_{\pm 2.36}$	$93.25_{\pm 3.36}$	$93.68_{\pm 3.83}$	$81.70_{\pm 11.33}$	$93.25_{\pm 3.36}$
	Full	$98.68_{\pm 0.71}$	$99.59_{\pm 0.32}$	$62.96_{\pm 22.43}$	$98.68_{\pm 0.71}$	$98.55_{\pm 0.64}$	$99.37_{\pm 0.42}$	$93.70_{\pm 9.92}$	$98.55_{\pm 0.64}$

Table 20: Test accuracy (%) on the four real tasks for GNN, GSAT, DIR, and Hybrid under Appendix C.8. Entries are means over thirty runs (three splits, ten seeds each), with sample standard deviations as subscripts. Every fitting budget is retained, including cases where GSAT or DIR exceeds Hybrid.

		GIN				GCN			
	t_n	GNN	GSAT	DIR	Hybrid	GNN	GSAT	DIR	Hybrid
AIDS	2	$63.80_{\pm 16.41}$	$65.13_{\pm 13.18}$	$84.32_{\pm 14.70}$	$63.80_{\pm 16.41}$	$59.10_{\pm 16.70}$	$64.63_{\pm 12.16}$	$61.30_{\pm 14.88}$	$59.10_{\pm 16.70}$
	5	$72.18_{\pm 12.69}$	$73.07_{\pm 8.85}$	$92.32_{\pm 7.78}$	$72.48_{\pm 12.86}$	$61.87_{\pm 13.08}$	$65.62_{\pm 11.99}$	$65.22_{\pm 12.93}$	$61.80_{\pm 13.10}$
	10	$76.65_{\pm 8.31}$	$75.95_{\pm 8.10}$	$91.65_{\pm 9.49}$	$78.23_{\pm 8.34}$	$65.97_{\pm 8.86}$	$69.13_{\pm 6.70}$	$65.32_{\pm 11.19}$	$68.15_{\pm 9.63}$
	20	$81.70_{\pm 7.74}$	$79.13_{\pm 7.07}$	$95.78_{\pm 5.41}$	$84.70_{\pm 7.78}$	$68.42_{\pm 7.58}$	$71.02_{\pm 5.68}$	$66.05_{\pm 9.78}$	$77.23_{\pm 6.81}$
	50	$86.93_{\pm 6.35}$	$81.68_{\pm 5.19}$	$96.52_{\pm 4.52}$	$90.90_{\pm 4.21}$	$73.07_{\pm 4.78}$	$72.78_{\pm 4.48}$	$69.02_{\pm 7.05}$	$81.20_{\pm 4.95}$
	100	$88.75_{\pm 4.97}$	$84.80_{\pm 4.86}$	$98.30_{\pm 2.78}$	$91.85_{\pm 3.78}$	$77.03_{\pm 3.22}$	$75.12_{\pm 3.20}$	$69.77_{\pm 8.71}$	$86.17_{\pm 3.32}$
	200	$93.80_{\pm 2.29}$	$93.27_{\pm 2.77}$	$98.40_{\pm 1.70}$	$95.17_{\pm 1.86}$	$78.90_{\pm 2.58}$	$76.80_{\pm 2.74}$	$74.37_{\pm 4.89}$	$87.07_{\pm 2.37}$
	Full	$98.08_{\pm 0.71}$	$98.62_{\pm 0.67}$	$97.42_{\pm 3.91}$	$98.27_{\pm 0.65}$	$88.72_{\pm 1.49}$	$87.38_{\pm 2.02}$	$79.18_{\pm 1.32}$	$90.42_{\pm 1.85}$
TOX21	2	$56.88_{\pm 5.16}$	$57.28_{\pm 3.92}$	$56.74_{\pm 4.27}$	$56.88_{\pm 5.16}$	$54.02_{\pm 4.58}$	$54.40_{\pm 4.63}$	$56.02_{\pm 5.22}$	$54.02_{\pm 4.58}$
	5	$56.84_{\pm 4.46}$	$57.68_{\pm 4.54}$	$56.60_{\pm 4.08}$	$58.25_{\pm 4.74}$	$55.42_{\pm 3.86}$	$56.19_{\pm 3.44}$	$56.49_{\pm 3.56}$	$57.18_{\pm 4.71}$
	10	$58.04_{\pm 4.04}$	$58.09_{\pm 3.32}$	$57.96_{\pm 4.71}$	$60.39_{\pm 5.30}$	$57.02_{\pm 3.51}$	$57.32_{\pm 3.68}$	$56.79_{\pm 5.07}$	$60.63_{\pm 5.11}$
	20	$59.79_{\pm 3.38}$	$58.98_{\pm 5.01}$	$59.61_{\pm 4.80}$	$63.95_{\pm 2.58}$	$58.77_{\pm 4.31}$	$59.04_{\pm 4.06}$	$59.86_{\pm 4.60}$	$63.16_{\pm 3.87}$
	50	$64.39_{\pm 3.50}$	$62.49_{\pm 3.73}$	$59.79_{\pm 3.15}$	$67.70_{\pm 2.55}$	$61.70_{\pm 3.88}$	$61.39_{\pm 4.25}$	$63.60_{\pm 4.10}$	$66.30_{\pm 3.51}$
	100	$66.89_{\pm 3.54}$	$65.58_{\pm 4.05}$	$61.42_{\pm 3.42}$	$69.81_{\pm 3.01}$	$63.70_{\pm 3.39}$	$63.07_{\pm 3.65}$	$63.49_{\pm 4.74}$	$69.56_{\pm 4.09}$
	200	$70.49_{\pm 3.07}$	$68.12_{\pm 3.28}$	$62.86_{\pm 3.96}$	$72.40_{\pm 2.81}$	$66.14_{\pm 4.91}$	$65.40_{\pm 4.11}$	$66.04_{\pm 3.26}$	$71.30_{\pm 4.65}$
	Full	$75.70_{\pm 1.73}$	$75.47_{\pm 2.12}$	$65.11_{\pm 3.33}$	$76.04_{\pm 1.72}$	$70.75_{\pm 2.89}$	$69.18_{\pm 2.70}$	$68.32_{\pm 4.37}$	$74.11_{\pm 3.34}$
DHFR	2	$60.18_{\pm 3.20}$	$59.87_{\pm 4.06}$	$60.31_{\pm 2.84}$	$60.18_{\pm 3.20}$	$59.33_{\pm 3.41}$	$59.29_{\pm 3.18}$	$55.11_{\pm 7.75}$	$59.33_{\pm 3.41}$
	5	$60.93_{\pm 2.60}$	$59.02_{\pm 4.29}$	$61.33_{\pm 2.27}$	$61.11_{\pm 3.36}$	$58.58_{\pm 5.49}$	$58.84_{\pm 4.03}$	$54.00_{\pm 8.10}$	$58.98_{\pm 5.24}$
	10	$59.33_{\pm 5.49}$	$58.98_{\pm 5.74}$	$59.87_{\pm 3.71}$	$59.91_{\pm 5.96}$	$60.58_{\pm 4.51}$	$58.58_{\pm 5.26}$	$56.71_{\pm 6.86}$	$61.29_{\pm 5.37}$
	20	$62.84_{\pm 6.20}$	$61.78_{\pm 5.62}$	$59.73_{\pm 4.18}$	$64.44_{\pm 7.17}$	$61.73_{\pm 5.77}$	$60.89_{\pm 4.89}$	$59.78_{\pm 7.41}$	$64.67_{\pm 5.74}$
	50	$69.78_{\pm 7.64}$	$65.56_{\pm 7.25}$	$60.22_{\pm 4.29}$	$71.38_{\pm 6.66}$	$68.89_{\pm 4.83}$	$64.71_{\pm 5.72}$	$63.47_{\pm 8.28}$	$70.00_{\pm 3.94}$
	100	$73.29_{\pm 5.30}$	$71.51_{\pm 5.45}$	$61.42_{\pm 3.45}$	$72.98_{\pm 5.68}$	$72.09_{\pm 5.55}$	$67.33_{\pm 4.44}$	$67.07_{\pm 7.66}$	$72.13_{\pm 6.01}$
	200	$78.31_{\pm 5.06}$	$74.44_{\pm 7.67}$	$61.07_{\pm 4.12}$	$78.53_{\pm 4.70}$	$74.76_{\pm 4.04}$	$71.51_{\pm 4.12}$	$66.04_{\pm 8.23}$	$75.82_{\pm 4.17}$
	Full	$80.40_{\pm 4.32}$	$76.00_{\pm 5.26}$	$61.47_{\pm 4.60}$	$80.62_{\pm 4.20}$	$75.78_{\pm 4.43}$	$72.09_{\pm 3.93}$	$71.91_{\pm 7.67}$	$76.89_{\pm 3.89}$
PROTEINS	2	$63.21_{\pm 4.54}$	$63.54_{\pm 5.96}$	$65.89_{\pm 7.03}$	$63.21_{\pm 4.54}$	$59.76_{\pm 8.78}$	$61.16_{\pm 7.81}$	$57.23_{\pm 8.34}$	$59.76_{\pm 8.78}$
	5	$63.30_{\pm 3.81}$	$63.30_{\pm 4.00}$	$68.04_{\pm 6.09}$	$63.78_{\pm 4.47}$	$61.40_{\pm 5.33}$	$62.53_{\pm 4.54}$	$56.49_{\pm 8.95}$	$62.95_{\pm 6.35}$
	10	$63.30_{\pm 4.79}$	$63.90_{\pm 4.71}$	$67.83_{\pm 4.88}$	$63.69_{\pm 4.55}$	$62.32_{\pm 5.31}$	$63.27_{\pm 4.55}$	$54.67_{\pm 9.55}$	$62.65_{\pm 7.02}$
	20	$64.08_{\pm 5.18}$	$64.88_{\pm 5.27}$	$68.81_{\pm 6.27}$	$65.00_{\pm 5.55}$	$63.12_{\pm 4.10}$	$63.57_{\pm 3.18}$	$58.45_{\pm 8.20}$	$64.73_{\pm 4.70}$
	50	$65.86_{\pm 4.36}$	$65.74_{\pm 4.79}$	$69.79_{\pm 5.67}$	$66.37_{\pm 5.23}$	$64.43_{\pm 5.12}$	$65.09_{\pm 4.28}$	$53.12_{\pm 10.93}$	$66.07_{\pm 5.78}$
	100	$66.99_{\pm 5.03}$	$67.23_{\pm 5.58}$	$69.17_{\pm 5.38}$	$67.83_{\pm 5.25}$	$65.48_{\pm 4.81}$	$65.68_{\pm 4.45}$	$49.73_{\pm 11.05}$	$66.85_{\pm 6.03}$
	200	$69.29_{\pm 5.29}$	$68.90_{\pm 5.07}$	$69.17_{\pm 5.69}$	$69.55_{\pm 5.43}$	$66.07_{\pm 4.59}$	$66.88_{\pm 4.73}$	$49.79_{\pm 11.59}$	$66.67_{\pm 5.09}$
	Full	$70.92_{\pm 4.81}$	$70.86_{\pm 4.80}$	$69.70_{\pm 6.97}$	$71.07_{\pm 4.57}$	$67.98_{\pm 4.20}$	$67.56_{\pm 3.48}$	$59.58_{\pm 9.46}$	$68.75_{\pm 4.79}$

E Pooled ERM on Graphs

This appendix reports the graph studies of Section 5 that train networks by pooled ERM. Their conditions, run order, and decision rules were fixed before any run, and every cell is reported. Each cell has thirty paired runs over three splits and ten seeds, or ten runs on the original MUTAGENICITY split. An asterisk marks a difference significant at the 5% level in an exact sign-flip test over the paired runs, stratified by split and Holm-corrected within each pre-registered family of cells. The networks follow the recipe of Appendix C.3 and ran with PyTorch 1.12.1 and PyTorch Geometric 2.3. Within a run, all networks share the initialization seed, so their differences come from the training objective.

Methods. The GNN is trained by cross-entropy on the available labels. Pooled ERM trains the same network on the weighted objective (8), with group majorities as targets and $|2\bar{Y}_t - 1|$ as weights, and selects its checkpoint by the validation accuracy of the pooled prediction. It predicts by averaging the trained network's hard predictions over training graphs with the query's key; a query whose key is unseen, or whose average is exactly one half, keeps the network's prediction. Selective pooled ERM pools only the groups that pass the gate of Appendix C.5 on the training labels, with the network's checkpoint accuracy as reference, and treats every other graph as its own group. The fixed-key hybrid votes on the same gated groups and falls back to the GNN, and the deployed hybrid is the procedure

of Table 2. A soft-label variant of pooled ERM, which fits the group label frequency by cross-entropy, differs from the weighted form on TOX21 by -0.74 to $+0.21$ points and never significantly.

Supplied keys. FOUR-MOTIF uses its sixteen planted motif sets as keys, and the explanation-class family uses the planted motif set of each graph. At noise level ρ, every fitting-pool and validation label is flipped independently with probability ρ, with one draw per split and seed shared across the nested budgets; test labels are never flipped, so test error is excess risk. Test-time voting replaces the GNN prediction by the unique fitting-label majority of the query's key, without support or purity gates.

Mined keys fixed before training. Under the strict budget of Appendix D.4, the candidate vocabulary is mined only from pool graphs outside the B budgeted labels, whose labels are never read, and from checkpoint graphs. Patterns are selected with checkpoint labels, using $(\tau, K) = (0.50, 10)$ and the checkpoint majority-class rate as reference. The key map is frozen before training and is therefore fixed given the training sample. One of the 480 maps selected no pattern; there every method equals the GNN.

E.1 Supplied Keys

Table 21 gives every FOUR-MOTIF setting and Table 22 the explanation-class family. Pooled ERM improves significantly over the GNN in all 94 settings. Under 20% noise its gain over test-time voting is positive for GCN and GIN at every budget up to fifty graphs per class, and five of these ten comparisons are significant. Without noise the two methods agree to within one point in nine of these ten comparisons, and at larger budgets both reach full accuracy. At $t_n = 20$ the gain over the GNN falls from about 40 points at $J = 4$ to about 20 points at $J = 56$.

Table 21: Pooled ERM on FOUR-MOTIF with supplied keys, for four backbones and two noise levels ρ. GNN is its test accuracy in percent; Pooled−GNN and Pooled−Vote are accuracy differences in points from the GNN and from test-time voting on the same keys. Fitting and validation labels are flipped with probability ρ, and test labels are clean. SAGE denotes GraphSAGE.

		GNN				Pooled−GNN				Pooled−Vote			
	t_n	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE
$\rho = 0.2$	2	50.65	54.93	55.41	51.21	$+9.67^*$	$+7.78^*$	$+7.68^*$	$+9.35^*$	$+1.28$	$+1.18$	$+1.09$	$+0.58$
	5	50.75	56.32	56.77	52.35	$+17.60^*$	$+13.76^*$	$+16.40^*$	$+17.66^*$	$+3.03^*$	$+2.34^*$	$+5.02^*$	$+2.66^*$
	10	52.83	55.88	57.73	54.30	$+24.94^*$	$+23.36^*$	$+24.32^*$	$+23.11^*$	$+2.49$	$+2.45$	$+4.08^*$	$+1.32$
	20	57.91	60.65	60.70	60.45	$+29.61^*$	$+29.72^*$	$+28.11^*$	$+28.28^*$	$+2.18$	$+3.63^*$	$+2.46^*$	$+2.95^*$
	50	67.11	63.66	63.41	67.80	$+30.32^*$	$+34.32^*$	$+33.63^*$	$+29.51^*$	$+3.19^*$	$+3.80^*$	$+2.36^*$	$+2.93^*$
	100	72.02	67.35	69.20	73.11	$+27.56^*$	$+32.55^*$	$+30.63^*$	$+26.64^*$	$+1.17^*$	$+1.32^*$	$+1.12^*$	$+1.17^*$
	200	80.30	76.91	75.17	78.58	$+19.67^*$	$+23.09^*$	$+24.83^*$	$+21.42^*$	-0.01	$+0.00$	$+0.01$	$+0.02$
	Full	95.09	94.80	93.09	95.67	$+4.91^*$	$+5.20^*$	$+6.91^*$	$+4.33^*$	$+0.00$	$+0.00$	$+0.00$	$+0.00$
$\rho = 0$	2	51.95	56.70	59.41	52.63	$+14.30^*$	$+11.98^*$	$+11.22^*$	$+14.16^*$	$+2.04^*$	$+0.62$	-0.12	$+0.18$
	5	52.76	58.66	60.76	54.66	$+25.60^*$	$+22.13^*$	$+21.85^*$	$+25.10^*$	-0.34	$+0.38$	$+1.29$	-0.01
	10	56.75	61.58	63.00	58.07	$+32.88^*$	$+29.50^*$	$+30.55^*$	$+32.12^*$	-0.08	$+0.65$	$+3.17^*$	$+0.38$
	20	64.92	63.86	66.63	67.56	$+32.53^*$	$+33.82^*$	$+32.44^*$	$+30.01^*$	$+0.38^*$	$+0.89^*$	$+1.75^*$	$+0.18$
	50	76.23	70.62	70.42	77.78	$+23.73^*$	$+29.31^*$	$+29.58^*$	$+22.19^*$	-0.01	$+0.02$	$+0.05$	$+0.05$
	100	85.41	81.04	74.85	83.09	$+14.59^*$	$+18.96^*$	$+25.15^*$	$+16.91^*$	$+0.00$	$+0.00$	$+0.00$	$+0.00$
	200	93.10	94.23	88.04	94.04	$+6.90^*$	$+5.77^*$	$+11.96^*$	$+5.96^*$	$+0.00$	$+0.00$	$+0.00$	$+0.00$
	Full	98.60	98.69	99.58	99.37	$+1.40^*$	$+1.31^*$	$+0.42^*$	$+0.63^*$	$+0.00$	$+0.00$	$+0.00$	$+0.00$

Table 22: Pooled ERM on the explanation-class family under 20% label noise, in the format of Table 21. The family with L motif types has $J = 1 + L + \binom{L}{2}$ keys, that is, $J = 4$, 22, and 56.

		GNN		Pooled−GNN		Pooled−Vote	
	t_n	GCN	GIN	GCN	GIN	GCN	GIN
$L = 2$	2	50.49	52.93	$+19.58^*$	$+19.68^*$	$+2.16$	$+1.49$
	5	51.58	53.60	$+37.20^*$	$+37.98^*$	-0.89	$+1.66$
	10	51.50	53.84	$+41.14^*$	$+40.06^*$	-0.57	$+0.48$
	20	57.23	54.97	$+39.91^*$	$+43.36^*$	$+1.66$	$+1.82$
	50	64.77	59.18	$+35.23^*$	$+40.82^*$	$+0.23$	$+0.32$
$L = 6$	2	50.03	52.25	$+8.08^*$	$+6.66^*$	$+1.68$	$+0.86$
	5	50.74	53.94	$+13.63^*$	$+12.16^*$	$+0.86$	$+1.52$
	10	51.73	56.85	$+21.09^*$	$+18.38^*$	$+1.66$	$+2.02$
	20	52.65	55.85	$+29.94^*$	$+29.31^*$	$+1.37$	$+3.25^*$
	50	57.44	57.80	$+35.54^*$	$+37.44^*$	$+1.05$	$+2.93^*$
$L = 10$	2	49.82	53.21	$+5.81^*$	$+4.36^*$	$+0.95$	-0.38
	5	50.09	54.08	$+12.22^*$	$+9.52^*$	$+1.33$	$+0.93$
	10	50.41	55.11	$+16.48^*$	$+13.64^*$	$+1.35$	$+1.38$
	20	52.42	56.48	$+22.12^*$	$+19.64^*$	$+2.00$	$+2.61$
	50	55.27	57.56	$+29.30^*$	$+27.41^*$	$+1.87$	$+1.22$

E.2 Mined Keys

Table 23 describes the key maps, and Tables 24 and 25 give pooled ERM and the two hybrids in all 64 settings. The maps select one to three patterns on average, giving two to ten keys among the training graphs and mean groups of 27 to 383 graphs. Nearly every test graph has a key seen in training, and the gated keys cover 11 to 48% of the test graphs. Pooling every group helps most at the smallest budget, significantly in nine of the sixteen settings at $B = 100$; at larger budgets the network separates graphs that a coarse key merges, which the approximation term of Theorem 3.1 prices. Selective pooling never lowers accuracy significantly and raises it in 24 settings, nine on TOX21, twelve on AIDS, one on DHFR, and two on PROTEINS. The fixed-key hybrid stays within 1.96 points of the deployed hybrid in every setting. Balanced accuracy follows the same pattern, and selective pooling gains up to 10 points on the imbalanced AIDS.

Table 23: Key maps mined only from graphs outside the training labels, averaged over the GCN and GIN runs. Pat. is the number of selected patterns, Keys the number of keys among the training graphs, Size the mean group size, Seen the share of test graphs whose key occurs in training, Gated the number of keys passing the gate, and Share the share of test graphs in gated keys.

	TOX21						AIDS					
B	Pat.	Keys	Size	Seen	Gated	Share	Pat.	Keys	Size	Seen	Gated	Share
100	1.20	2.3	36.9	0.999	0.73	0.204	1.00	4.7	38.0	0.967	0.77	0.429
200	1.70	3.4	59.1	0.999	0.95	0.179	1.37	2.7	66.9	1.000	0.97	0.478
500	1.50	3.0	178.0	1.000	0.83	0.178	1.40	2.9	181.9	1.000	0.98	0.473
1000	1.57	3.3	383.3	1.000	0.62	0.112	1.47	2.9	374.7	1.000	0.78	0.386
	DHFR						PROTEINS					
B	Pat.	Keys	Size	Seen	Gated	Share	Pat.	Keys	Size	Seen	Gated	Share
100	1.50	2.8	31.6	1.000	0.82	0.173	1.67	3.4	27.4	1.000	0.58	0.166
200	1.90	3.9	51.0	0.998	1.13	0.187	2.37	5.5	37.4	0.997	1.40	0.217
500	2.30	5.1	124.3	0.999	1.45	0.160	2.70	7.3	93.5	1.000	1.73	0.209
1000	2.17	4.8	175.2	0.999	1.30	0.118	3.07	9.8	147.9	0.999	1.97	0.179

Table 24: Pooled ERM with mined keys under the strict budget. GNN is the test accuracy in percent of the network trained on the $B - v$ training labels; Sel.−GNN and Pooled−GNN are accuracy differences in points for selective pooled ERM and for pooled ERM on every group. SAGE denotes GraphSAGE.

		GNN				Sel.−GNN				Pooled−GNN			
	B	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE
TOX21	100	59.28	60.14	63.33	59.56	$+2.61^*$	$+2.40^*$	$+2.02$	$+2.56^*$	$+5.40^*$	$+5.00^*$	$+1.42$	$+5.12^*$
	200	61.63	63.75	65.81	63.84	$+2.37^*$	$+2.33^*$	$+2.40^*$	$+1.54$	$+1.68$	-0.54	-2.09	-0.54
	500	65.51	69.75	71.32	69.35	$+2.25^*$	$+1.72^*$	$+0.00$	$+0.81$	-1.95	-5.95^*	-7.75^*	-5.84^*
	1000	67.56	72.56	74.65	71.32	$+1.65^*$	$+0.51$	$+0.61$	-0.25	-3.05^*	-8.09^*	-10.11^*	-6.81^*
AIDS	100	77.97	83.33	77.97	76.95	$+1.88^*$	$+0.90$	$+3.17^*$	$+3.17^*$	$+2.15^*$	-3.80^*	$+1.57$	$+3.10^*$
	200	77.98	86.08	80.93	78.58	$+3.83^*$	$+2.90^*$	$+5.05^*$	$+4.98^*$	$+2.42^*$	-5.47^*	-0.43	$+1.73^*$
	500	81.23	91.55	88.48	83.27	$+4.40^*$	$+1.48$	$+2.95^*$	$+4.92^*$	-0.73	-10.73^*	-7.57^*	-2.50^*
	1000	85.27	93.95	92.50	87.48	$+2.65^*$	$+0.72$	-0.17	$+2.23^*$	-3.83^*	-12.40^*	-11.15^*	-5.88^*
DHFR	100	62.98	65.11	68.40	64.67	$+0.76$	$+0.89$	-0.93	$+1.64$	-0.40	-2.53	-6.22^*	-2.18
	200	69.73	69.02	75.60	68.71	$+0.04$	$+1.07$	-0.00	$+0.67$	-6.13^*	-5.69^*	-12.27^*	-4.98^*
	500	74.31	78.36	81.24	73.91	$+1.38$	-0.00	-0.80	$+1.16$	-9.33^*	-13.82^*	-16.80^*	-9.69^*
	1000	75.87	80.13	82.27	74.84	$+0.93$	$+1.07$	-0.13	$+3.29^*$	-11.60^*	-15.91^*	-18.40^*	-10.44^*
PROTEINS	100	63.93	64.26	62.62	64.35	$+3.18^*$	$+0.36$	$+1.64$	$+1.85$	$+4.70^*$	$+3.66^*$	$+6.10^*$	$+4.37^*$
	200	64.61	67.08	64.32	64.70	$+1.85^*$	$+0.09$	$+1.58$	$+1.19$	$+1.34$	-1.46	$+1.13$	$+1.70$
	500	66.07	68.96	67.62	65.65	$+1.13$	$+1.28$	$+0.62$	$+0.57$	$+2.44$	-1.19	$+0.98$	$+2.98^*$
	1000	66.90	70.89	68.24	66.40	$+1.01$	-0.36	-1.07	$+0.68$	$+0.63$	-3.07^*	-1.16	$+0.65$

Table 25: The fixed-key and deployed hybrids with mined keys under the strict budget, as accuracy differences in points, with rows as in Table 24. Fixed votes with the key map mined outside the training labels and Dep. is the procedure of Table 2. The networks are retrained for this study, so Dep.−GNN differs slightly from Table 13.

		Fixed−GNN				Dep.−GNN				Fixed−Dep.			
	B	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE	GCN	GIN	PNA	SAGE
TOX21	100	$+3.28^*$	$+2.68^*$	$+2.04^*$	$+3.16^*$	$+2.63^*$	$+2.26^*$	$+2.09^*$	$+3.30^*$	$+0.65$	$+0.42$	-0.05	-0.14
	200	$+3.35^*$	$+1.84^*$	$+1.07^*$	$+1.70^*$	$+3.79^*$	$+1.82^*$	$+1.16^*$	$+2.74^*$	-0.44	$+0.02$	-0.09	-1.04
	500	$+2.86^*$	$+1.14^*$	$+0.18$	$+1.86^*$	$+2.79^*$	$+0.96^*$	$+0.58^*$	$+1.88^*$	$+0.07$	$+0.18$	-0.40	-0.02
	1000	$+2.05^*$	$+0.28$	$+0.04$	$+0.86^*$	$+2.75^*$	$+0.56^*$	$+0.16$	$+1.46^*$	-0.70	-0.28	-0.12	-0.60
AIDS	100	$+1.42^*$	$+0.95$	$+2.73^*$	$+2.80^*$	$+0.95^*$	$+0.35$	$+1.87^*$	$+2.22^*$	$+0.47$	$+0.60$	$+0.87$	$+0.58$
	200	$+3.12^*$	$+2.65^*$	$+3.43^*$	$+3.50^*$	$+3.00^*$	$+2.02^*$	$+2.97^*$	$+3.27^*$	$+0.12$	$+0.63^*$	$+0.47$	$+0.23$
	500	$+3.25^*$	$+1.65^*$	$+3.08^*$	$+3.37^*$	$+3.05^*$	$+0.83^*$	$+1.27^*$	$+2.85^*$	$+0.20$	$+0.82$	$+1.82^*$	$+0.52$
	1000	$+2.17^*$	$+0.58^*$	$+0.52^*$	$+1.85^*$	$+1.97^*$	$+0.32$	$+0.20$	$+1.73^*$	$+0.20$	$+0.27$	$+0.32$	$+0.12$
DHFR	100	$+0.62$	$+1.82^*$	$+0.84^*$	$+1.91$	$+0.09$	$+1.33$	$+0.36$	$+1.02$	$+0.53$	$+0.49$	$+0.49$	$+0.89$
	200	$+0.62$	$+1.60^*$	$+0.49$	$+1.16$	$+1.07$	$+1.64^*$	$+0.27$	$+1.02$	-0.44	-0.04	$+0.22$	$+0.13$
	500	$+1.29$	$+0.62$	-0.04	$+1.11^*$	$+1.29$	$+0.49$	$+0.09$	$+1.42^*$	$+0.00$	$+0.13$	-0.13	-0.31
	1000	$+0.44$	-0.09	-0.31	$+1.33^*$	$+0.27$	$+0.22$	$+0.00$	$+0.80$	$+0.18$	-0.31	-0.31	$+0.53$
PROTEINS	100	$+2.86^*$	$+0.95$	$+1.76$	$+2.32^*$	$+0.89$	$+0.54$	$+1.49$	$+1.43$	$+1.96$	$+0.42$	$+0.27$	$+0.89$
	200	$+1.25$	$+0.57$	$+1.07$	$+1.31^*$	$+1.37$	$+0.42$	$+1.16$	$+0.77$	-0.12	$+0.15$	-0.09	$+0.54$
	500	$+1.13$	$+0.68$	$+0.30$	$+0.89$	$+1.61^*$	$+0.71^*$	$+0.21$	$+1.13$	-0.48	-0.03	$+0.09$	-0.24
	1000	$+0.71$	$+0.12$	$+0.30$	$+0.83$	$+0.45$	$+0.03$	$+0.24$	$+0.18$	$+0.27$	$+0.09$	$+0.06$	$+0.65$

E.3 Supplied Descriptors on MUTAGENICITY

Keys are the sets of the ten supplied descriptors of Appendix C.4, and Table 26 gives every budget. Pooled ERM is ahead of the GNN in twelve of the fourteen finite budgets up to 200 per class, significantly for GIN at $t_n = 5$ after Holm correction over these fourteen cells. With the full pool the network alone separates graphs that share a descriptor set, so supplied descriptors help most when labels are scarce; test-time voting on the same keys behaves the same way.

Table 26: Pooled ERM with the supplied MUTAGENICITY descriptors on the original split, with ten paired runs per cell. GNN and Pooled are test accuracies in percent, and the differences are in points, with wins and losses for Pooled−GNN. Vote is test-time voting on the same keys.

		GNN		Pooled		Pooled−GNN (W/L)		Vote−GNN		Pooled−Vote	
	t_n	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN	GCN	GIN
MUTAGENICITY	2	63.44	63.14	65.80	64.76	+2.36 (6/2)	+1.62 (7/1)	+4.71	+3.70	−2.36	−2.08
	5	63.30	63.95	68.01	68.55	+4.71 (8/1)	+4.60* (10/0)	+6.63*	+4.41*	−1.92	+0.18
	10	66.19	66.10	70.74	69.91	+4.55 (9/1)	+3.81 (9/0)	+1.45	+1.02	+3.09	+2.79
	20	65.80	67.34	68.85	70.44	+3.05 (8/2)	+3.09 (8/1)	+3.23	+2.29	−0.18	+0.81
	50	68.66	69.58	70.16	71.18	+1.50 (8/2)	+1.59 (7/3)	+1.99	+0.67	−0.48	+0.92
	100	71.41	70.72	71.22	73.09	−0.18 (5/5)	+2.38 (8/2)	−0.28	+0.60	+0.09	+1.78
	200	73.46	72.73	72.89	73.03	−0.58 (2/7)	+0.30 (6/4)	−0.35	+0.25	−0.23	+0.05
	Full	79.49	79.65	73.12	73.42	−6.37* (0/10)	−6.24* (0/10)	−4.90*	−5.06*	−1.48*	−1.18*