CALIBHYPER: CHANCE-CORRECTED RELATIONAL HYPERGRAPHS FOR FEW-SHOT MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular property prediction is central to drug development and materials discovery, but experiments are costly and labeled data are scarce. Context-aware methods use auxiliary assay labels to support few-shot prediction, and recent work supervises property relations with label agreement. However, label agreement is sensitive to class marginals and does not directly capture dependence between properties. We propose CalibHyper, a chance-corrected relational hypergraph method based on the joint label distribution. CalibHyper subtracts an independence baseline from the ordered four-state label distribution and shrinks the residual according to the number of joint observations. A swap-equivariant relation head estimates these residuals, which choose the auxiliary properties for each molecule and set the sign and weight of their hyperedge messages. On thirteen datasets from five benchmarks, in both 1-shot and 10-shot settings, CalibHyper and its ablation settings achieve ROC-AUC competitive with the strongest reported results.[1]

1 INTRODUCTION

Molecular property prediction is a fundamental task in drug development and materials discovery (Wu et al., 2018; David et al., 2020), but experimental labels are expensive to obtain (Wang et al., 2025c). Few-shot methods reuse knowledge learned on existing properties and adapt to an unseen property from a small support set (Altae-Tran et al., 2017; Guo et al., 2021; Chen et al., 2023; Wu et al., 2024; Li et al., 2025). Labels from existing auxiliary assays can also serve as context for prediction (Figure 1A). GS-Meta (Zhuang et al., 2023) and KRGTS (Wang et al., 2025b) organize this information with graphs that connect molecules to properties and with relations among tasks, while methods for auxiliary data selection and task weighting consider how well the auxiliary information matches the target task (Huang et al., 2023; Zhong & Mottin, 2025).

ReCoG (Wang et al., 2026) supervises relation learning with binary label agreement, which mixes the dependence between two properties with their class marginals. When positives are rare, two independent properties still agree on almost every molecule, because most molecules are negative for both (Figure 1B): MUV-713 and MUV-733 agree on 99.8718% of their 3,121 jointly labeled molecules, against 99.8719% expected under independence (Cohen's $\kappa = -0.0005$). Relation supervision should therefore separate chance agreement, which the marginals explain, from dependence beyond it.

Binary agreement also merges the two ordered disagreement states, target positive with auxiliary negative and the reverse (Figure 1C). The association itself can change with molecular structure (Figure 1D), and an estimate from few joint observations is less reliable than one from many (Figure 1E). In ReCoG, the relation head affects prediction only through an auxiliary loss on shared representations, and a gate weights each auxiliary property once per episode. We instead let relation estimates drive messages that differ from molecule to molecule.

We propose CalibHyper, a chance-corrected relational hypergraph method whose two modules share one relation estimate. C4RC (chance-corrected four-state relation calibration) keeps the four ordered

[1] Code and data: `https://anonymous.4open.science/r/CalibHyper-0FE1/`.

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

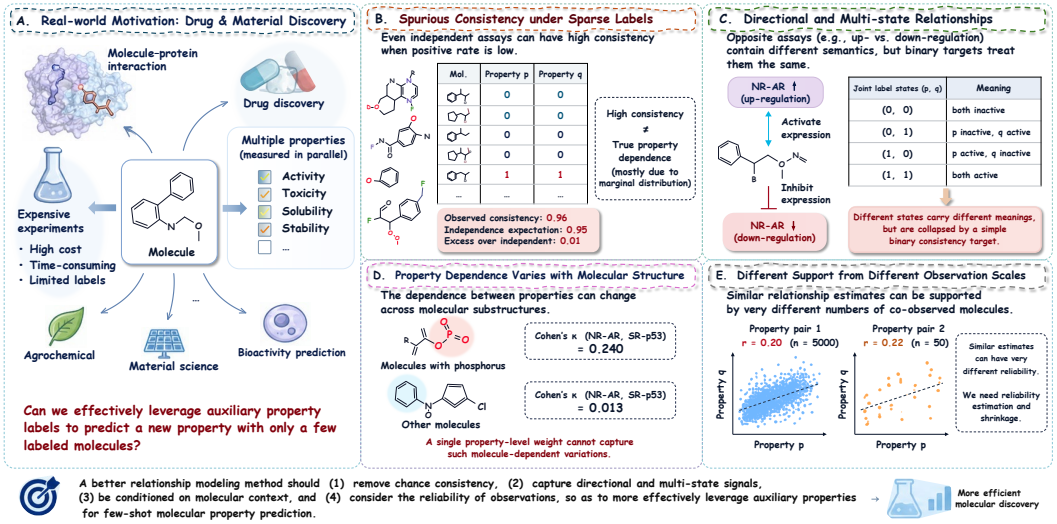


Figure 1: Motivation. (A) Few-shot prediction with auxiliary assays. (B) Chance agreement under rare positives. (C) Binary targets merge states 01 and 10. (D) Structure-dependent association in Tox21. (E) Similar estimates from unequal numbers of joint observations. B, C, and E are schematic.

label states, subtracts the independence baseline computed on jointly observed molecules, shrinks the residual according to the number of joint observations, and fits it with a swap-equivariant head. MSRHA (molecule-conditioned signed relational hypergraph adapter) uses the predicted residuals to select auxiliary properties for each molecule and to set the sign and weight of messages along ordered hyperedges of a molecule, the target, and an auxiliary property.

Our contributions are as follows.

- Using real molecular label matrices, we show that binary agreement supervision confounds class marginals with property dependence and that swap-invariant relation inputs cannot represent ordered label states.
- We propose CalibHyper, which couples four-state relation calibration with signed hypergraph propagation so that relation estimates drive molecule-conditioned auxiliary messages. We characterize the calibrated target and prove that the relation head is swap-equivariant and that the adapter is the identity at initialization.
- On 26 conditions (thirteen datasets, 1-shot and 10-shot), CalibHyper and its ablation settings achieve ROC-AUC competitive with the strongest reported results, with larger margins in the 1-shot setting. Ablations and interventions on auxiliary labels and routing show how each component and the available context affect performance.

2 Related Work

Few-shot and context-aware molecular property prediction. Few-shot methods transfer supervision from existing properties to new ones through shared representations, sample matching, and task adaptation (Wang et al., 2025c), as in PAR (Wang et al., 2021), ADKF-IFT (Chen et al., 2023), PACIA (Wu et al., 2024), Pin-Tuning (Wang et al., 2024), and UniMatch (Li et al., 2025). Multi-task models share statistical strength across the property labels of the same molecule (Ramsundar et al., 2015; Mayr et al., 2018), and context-aware methods extend this idea to unseen properties by exploiting context molecules, measurements, or context graphs, as in MHNfs (Schimunek et al., 2023), CAMP (Fifty et al., 2023), GS-Meta (Zhuang et al., 2023), KRGTS (Wang et al., 2025b), and CaMol (Hoang & Lee, 2026); graph prompting methods such as OFA (Liu et al., 2024) organize tasks with prompt graphs. The closest work, ReCoG (Wang et al., 2026), supervises property relations with binary agreement and compresses auxiliary property nodes conditioned on the target.

CalibHyper replaces binary agreement with chance-corrected four-state residuals and uses them to generate molecule-conditioned messages.

Selecting and compressing auxiliary information. MolGroup (Huang et al., 2023) and AutAuT (Zhong & Mottin, 2025) select or weight auxiliary information at the dataset and task level. Information bottleneck methods such as CGIB (Lee et al., 2023), IGIB-ISE (Zhang et al., 2025), and InfoAlign (Liu et al., 2025a) constrain what a representation retains. We combine auxiliary-node compression with a chance-corrected relation target.

Chance-corrected agreement and higher-order relations. Cohen's κ (Cohen, 1960) separates observed agreement from the chance agreement implied by the marginals, and under skewed marginals high agreement can coexist with low κ (Feinstein & Cicchetti, 1990; Byrt et al., 1993); we carry this correction into relation supervision as a four-state residual. Hypergraph learning (Feng et al., 2019; Chien et al., 2022; Feng et al., 2025; Zhang et al., 2026; Gao et al., 2026), signed networks (Derr et al., 2018), and dynamic message passing (Zhang et al., 2020) handle higher-order connections, signed relations, and input-dependent neighborhoods. Our hyperedges are ordered triples of a molecule, the target, and an auxiliary property, and one supervised relation score sets their selection, weight, and sign. Appendix A gives an extended review.

3 Problem Setup and Notation

Let $\mathcal{M}$ be a set of molecules and $\mathcal{T}$ a set of binary property prediction tasks, split into disjoint meta-training and meta-test sets, $\mathcal{T}_{\text{train}} \cap \mathcal{T}_{\text{test}} = \emptyset$. The label matrix is $Y \in \{0, 1, \star\}^{|\mathcal{M}| \times |\mathcal{T}|}$, where $y_{i,p} = 1$ means that molecule m_i is active for property p, 0 means inactive, and $\star$ means unmeasured. We write $o_{i,p} = \mathbf{1}[y_{i,p} \neq \star]$ for the observation indicator.

An episode $E_\tau = (S_\tau, Q_\tau)$ targets property τ. The support set S_τ usually contains K molecules per class, and the query set Q_τ contains M molecules. The auxiliary set is $A_\tau \subset \mathcal{T}_{\text{train}} \setminus \{\tau\}$ with $N = |A_\tau|$. The context graph G_τ has molecule nodes V_τ^m from the episode and property nodes $V_\tau^p = \{\tau\} \cup A_\tau$; observed labels determine the edge types, and query labels of the target are excluded. Molecule features come from a pretrained GIN (Xu et al., 2019; Hu et al., 2020), $x_i = F_{\text{mol}}(m_i) \in \mathbb{R}^d$, and property nodes use learnable embeddings. A two-layer context encoder produces $H = \{h_v\}_{v \in V_\tau} = \text{ContextEncoder}(G_\tau, X)$, and the predictor outputs $\hat{y}_{i,\tau} = F_{\text{pred}}([\, h'_i \| h'_\tau \,])$, where H' are the representations after adaptation ($H' = H$ without an adapter). A context information bottleneck compresses auxiliary property representations conditioned on the target (Lee et al., 2023; Wang et al., 2026) and assigns each auxiliary property a gate $\lambda_a \in (0, 1)$ through a Gumbel-sigmoid relaxation (Jang et al., 2017; Maddison et al., 2017). Training uses MAML-style bilevel optimization (Finn et al., 2017): the inner loop adapts on the support loss, and the outer loop updates shared parameters with the query loss and auxiliary objectives.

The joint label of a property pair (p, q) is one-hot encoded as $e(u, v) \in \{0, 1\}^4$ with coordinates $(e_{00}, e_{01}, e_{10}, e_{11})$, where the first digit refers to p. The appendix gives the notation, formal definitions, Propositions 2 to 5, the proofs of all propositions, and further analysis.

4 Limits of Binary Relation Supervision

ReCoG (Wang et al., 2026) regresses the disagreement indicator $y_{\text{rel}}^{(i,p,q)} = (y_{i,p} - y_{i,q})^2 \in \{0, 1\}$ of each property pair with an MSE loss.

Chance agreement. Let π_p and π_q be the positive rates of p and q on their jointly observed molecules, a_{pq} their observed agreement, and $c_{pq} = (1 - \pi_p)(1 - \pi_q) + \pi_p \pi_q$ the agreement expected under independence.

Proposition 1 (Chance decomposition)**.** *If $c_{pq} < 1$, then $\mathbb{E}[y_{rel}] = 1 - a_{pq} = (1 - c_{pq})(1 - \kappa_{pq})$, where $\kappa_{pq} = (a_{pq} - c_{pq})/(1 - c_{pq})$ is Cohen's κ. With $s_{i,p} = 1 - 2y_{i,p}$, the four-state indicators satisfy*

$$(e_{00} + e_{11}) - (e_{01} + e_{10}) = s_{i,p}\, s_{i,q}, \tag{1}$$

whose mean exceeds its independence value $(1 - 2\pi_p)(1 - 2\pi_q)$ by $4\,\text{Cov}(y_p, y_q) = 2(a_{pq} - c_{pq})$.

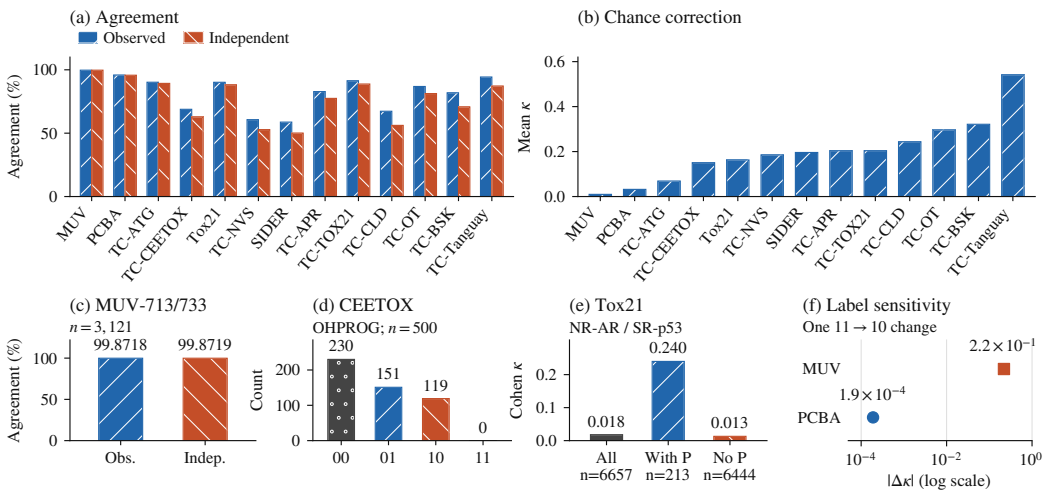


Figure 2: Label statistics (Appendix H). (a) Observed and independence agreement and (b) mean Cohen's κ per dataset (TC: ToxCast). (c) MUV-713 and MUV-733. (d) A CEETOX increase and decrease pair. (e) κ with and without phosphorus. (f) $|\Delta\kappa|$ after one $11 \to 10$ relabeling.

When positives are rare, $c_{pq} \approx 1$, so dependence moves the binary target only in proportion to $1 - c_{pq}$. For MUV-713 and MUV-733, $1 - c_{pq} = 0.0013$, so each unit of κ_{pq} changes the expected binary target by only 0.0013. Figure 2 shows that this regime is common. The mean κ is 0.010 on MUV and 0.032 on PCBA, where joint negatives make up on average more than 99.99% and 99.8% of agreeing label pairs. Negative dependence is also frequent: 20.4%, 8.7%, and 15.6% of meta-training pairs have $\kappa < -0.02$ in the ToxCast subsets NVS, BSK, and CEETOX. Evidence is uneven as well. The median number of joint observations is 235,660 on PCBA but 22 on NVS, and two pairs with $\hat{\kappa} \approx 0.22$ differ by about 1,200-fold in their sensitivity to a single relabeled molecule (Figure 2f).

Swap-invariant relation input. ReCoG's relation head reads $u = [h_i \| h_p] \odot [h_i \| h_q] = [h_i \odot h_i \| h_p \odot h_q]$. The first half ignores the property pair, the second half ignores the molecule, and u is unchanged when p and q are swapped. A swap-equivariant four-state output computed from u must therefore assign equal values to the states 01 and 10, and its squared error on an ordered target is at least $\frac{1}{2}(r_{01} - r_{10})^2$ (Proposition 2). The two states do differ in practice: for the CEETOX assays in Figure 2d they contain 151 and 119 molecules, which the binary label merges.

5 Method

CalibHyper has two modules (Figure 3). C4RC learns reliability-weighted four-state residuals with a swap-equivariant head F_{rel} (Section 5.1), and MSRHA turns them into molecule-conditioned signed hyperedge messages that update the representations used for prediction (Section 5.2). The backbone follows Section 3.

5.1 C4RC: Chance-Corrected Four-State Relation Calibration

Independence baseline. For an ordered pair (p, q) of meta-training properties, let $I_{pq} = \{i : o_{i,p} = o_{i,q} = 1\}$, $n_{pq} = |I_{pq}|$, and $n^{+}_{p|q} = \sum_{i \in I_{pq}} y_{i,p}$. With a smoothing constant α (1 by default), the positive rate of p on the jointly observed molecules is $\hat{\pi}_{p|q} = (n^{+}_{p|q} + \alpha)/(n_{pq} + 2\alpha)$, and $\hat{\pi}_{q|p}$ is defined symmetrically. Two independent properties with these marginals have the four-state distribution

$$b_{pq} = \Big[(1 - \hat{\pi}_{p|q})(1 - \hat{\pi}_{q|p}),\ (1 - \hat{\pi}_{p|q})\hat{\pi}_{q|p},\ \hat{\pi}_{p|q}(1 - \hat{\pi}_{q|p}),\ \hat{\pi}_{p|q}\hat{\pi}_{q|p}\Big] \in \Delta^3.$$

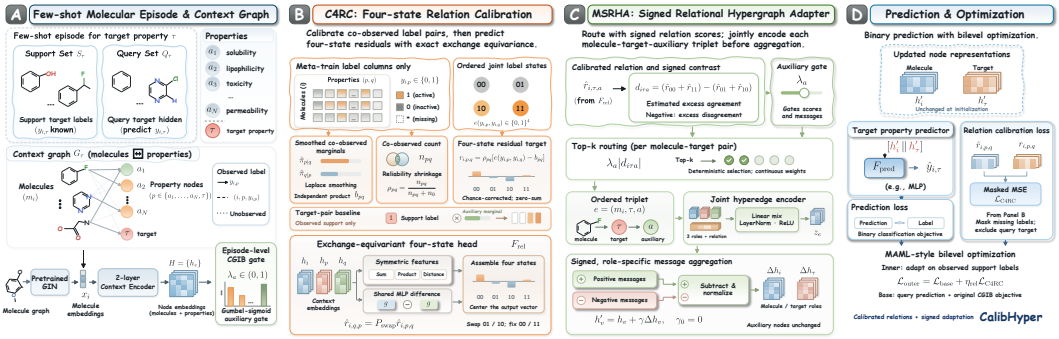


Figure 3: Architecture. (A) Context graph and encoder. (B) C4RC: chance-corrected four-state relation head. (C) MSRHA: signed messages on top-k hyperedges. (D) Prediction and training.

Reliability and target. The reliability $\rho_{pq} = n_{pq}/(n_{pq} + n_0)$ with $n_0 > 0$ shrinks the target toward zero when joint observations are few; with $n_0 = 5$, the NVS median gives $\rho_{pq} = 0.815$ and the PCBA median 0.99998. Under a signal and noise model given in the appendix, it is the linear shrinkage with minimum mean squared error when n_0 equals the noise-to-signal ratio (Robbins, 1956; James & Stein, 1961); we treat n_0 as a hyperparameter. For a triple (i, p, q) with $p \neq q$ and both labels observed,

$$r_{i,p,q} = \rho_{pq}\Big[\, e(y_{i,p}, y_{i,q}) - b_{pq} \,\Big] \in \mathbb{R}^4, \qquad \textstyle\sum_k r_{i,p,q}[k] = 0. \tag{2}$$

With exact marginals, $\mathbb{E}[r_{i,p,q}] = \mathbf{0}$ when p and q are independent (Proposition 3). The signed contrast of the target and the signed score $1 - 2\hat{y}_{\text{rel}}$ of a binary head that fits $\mathbb{E}[y_{\text{rel}}]$ have means

$$\mathbb{E}\big[(r_{00} + r_{11}) - (r_{01} + r_{10})\big] = 2\rho_{pq}(1 - c_{pq})\kappa_{pq}, \qquad \bar{d}^{\,\text{bin}}_{pq} = (2c_{pq} - 1) + 2(1 - c_{pq})\kappa_{pq}, \tag{3}$$

so only the binary score carries the marginal term $2c_{pq} - 1$.

Unseen target properties. Support sets are stratified by label, so the positive fraction of the target in a support set reflects the sampling rule rather than its marginal. For pairs that involve an unseen target τ, we condition on the support label and subtract only the auxiliary marginal:

$$b_{i,\tau q} = \Big[(1 - y_{i,\tau})(1 - \hat{\pi}_q),\ (1 - y_{i,\tau})\hat{\pi}_q,\ y_{i,\tau}(1 - \hat{\pi}_q),\ y_{i,\tau}\hat{\pi}_q\Big], \qquad \rho_{\tau q} = \frac{n_q}{n_q + n_0},$$

where n_q and $\hat{\pi}_q$ are the number of observed labels and the smoothed positive rate of q on the meta-training columns. Query labels of the target are never used for relation targets, graph construction, or routing.

Swap-equivariant head. Swapping the two properties should exchange the two disagreement coordinates and leave the other two unchanged. We therefore combine the swap-invariant features

$$\xi_{ipq} = \big[\, h_i \odot (h_p + h_q) \,\|\, h_p \odot h_q \,\|\, |h_p - h_q| \,\big] \in \mathbb{R}^{3d}$$

with the antisymmetric scalar $\delta_{ipq} = g(h_i \odot h_p) - g(h_i \odot h_q) = -\delta_{iqp}$ from an MLP $g : \mathbb{R}^d \to \mathbb{R}$. An MLP $f : \mathbb{R}^{3d} \to \mathbb{R}^3$ maps ξ_{ipq} to $(\hat{r}_{00}, \hat{r}_{11}, u_c)$, where u_c is the sum of the two disagreement coordinates, and δ_{ipq} splits it:

$$\hat{r}_{01} = \tfrac{1}{2}(u_c + \delta_{ipq}), \qquad \hat{r}_{10} = \tfrac{1}{2}(u_c - \delta_{ipq}). \tag{4}$$

After centering, $\hat{r} \leftarrow \hat{r} - \frac{1}{4}\sum_k \hat{r}[k]$, the head satisfies $\hat{r}_{iqp} = P_{\text{swap}}\hat{r}_{ipq}$ exactly, where P_{swap} exchanges 01 and 10 (Proposition 4). Unlike the input of ReCoG, ξ_{ipq} contains cross terms between molecule and property, and δ_{ipq} keeps the direction of the relation. The head is trained with the masked loss $\mathcal{L}_{\text{C4RC}} = \sum m_{i,p,q}\|\hat{r}_{i,p,q} - r_{i,p,q}\|_2^2 / \big(4\sum m_{i,p,q}\big)$ over valid triples. The outer objective is $\mathcal{L}_{\text{outer}} = \mathcal{L}_{\text{base}} + \eta_{\text{rel}}\mathcal{L}_{\text{C4RC}}$, where $\mathcal{L}_{\text{base}}$ collects the remaining terms of the ReCoG objective, including the query prediction loss and the bottleneck objective.

5.2 MSRHA: Molecule-Conditioned Signed Relational Hypergraph Adapter

Signed agreement. For a molecule i, the episode target τ, and an auxiliary property $a \in A_\tau$, we evaluate the head on the triple and take the signed contrast

$$d_{i\tau a} = \big(\hat{r}_{00} + \hat{r}_{11}\big) - \big(\hat{r}_{01} + \hat{r}_{10}\big) \in \mathbb{R}. \tag{5}$$

The sign of $d_{i\tau a}$ selects the message channel and its magnitude sets the weight.

Molecule-conditioned sparse routing. For each molecule and target, we keep the k auxiliary properties with the highest gated scores:

$$\beta_{i\tau a} = \lambda_a \big|d_{i\tau a}\big|, \qquad \mathcal{A}_{i\tau} = \text{Top-}k_{\,a \in A_\tau}\ \beta_{i\tau a}.$$

Training passes with the bottleneck use λ_a; ordinary prediction and testing set $\lambda_a = 1$.

Ordered hyperedges. For each $a \in \mathcal{A}_{i\tau}$, the ordered triple $e = (m_i, \tau, a)$ is encoded jointly:

$$z_e = \mathrm{ReLU}\Big(\mathrm{LN}\big(W_m h_i + W_t h_\tau + W_a h_a + W_r \hat{r}_{i\tau a} + b\big)\Big) \in \mathbb{R}^{d_h}.$$

Unlike a sum of pairwise terms $\psi_{12} + \psi_{13} + \psi_{23}$, this encoder can express three-way interactions.

Signed, role-aware write-back. Each edge carries an agreement weight $w_e^+ = \lambda_a \,\mathrm{ReLU}(d_{i\tau a})$ and an opposition weight $w_e^- = \lambda_a \,\mathrm{ReLU}(-d_{i\tau a})$, and messages are aggregated with projections specific to each sign and node role $\mathrm{rl}(v) \in \{\text{molecule}, \text{target}\}$:

$$\Delta h_v = \frac{1}{Z_v} \sum_{e \ni v} \Big(w_e^+ T^+_{\mathrm{rl}(v)}(z_e) - w_e^- T^-_{\mathrm{rl}(v)}(z_e)\Big), \qquad Z_v = \varepsilon + \sum_{e \ni v} \big(w_e^+ + w_e^-\big), \tag{6}$$

where $T^\pm_{\mathrm{rl}}$ are role-specific affine maps. Dividing by Z_v makes Δh_v a weighted average of edge messages, so its scale does not grow with the number of retained edges. Only molecule and target nodes are updated.

Zero-initialized residual gate. The update enters as $h'_v = h_v + \gamma\, \Delta h_v$ with $\gamma_0 = 0$. Because LayerNorm (Ba et al., 2016) sits inside the message, a zero gate reproduces the predictions and the prediction-loss gradients of the model without the adapter, while γ itself still receives a gradient (Proposition 5). Depending on the ablation setting, γ is learned or frozen at zero. The adapter scores N candidate triples per molecule and passes messages along $\min(k, N)$ of them, so its cost grows linearly with N. On Tox21, C4RC and MSRHA add 588K parameters (19.50%); Appendix D gives the ablation settings and the complexity.

6 Experiments

6.1 Setup

Data and tasks. We use Tox21, SIDER, MUV, ToxCast, and PCBA from MoleculeNet (Wu et al., 2018) with the property splits of GS-Meta (Zhuang et al., 2023). Evaluating the nine ToxCast assay providers separately gives thirteen datasets and 26 conditions in the 1-shot and 10-shot settings. Each episode samples K support molecules per class, and the remaining observed molecules of the target form the query set. In the default warm setting, query molecules keep their auxiliary labels on meta-training properties.

Model and training. The encoder is a five-layer pretrained GIN with 300 hidden dimensions, followed by the two-layer context encoder. Training usually runs for 2,000 episodes with evaluation every 100 episodes and query batches of 16 for training and 64 for testing. Appendix C.1 gives the splits, sampling, evaluation protocol, and hyperparameters.

Table 1: ROC-AUC (%). CalibHyper: mean ± sample standard deviation over three seeds; baselines from Wang et al. (2026). ToxCast averages its nine subsets; a dash marks an unreported result.

	Tox21		SIDER		MUV		ToxCast		PCBA	
Method	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot	10-shot	1-shot
Siamese	80.40±0.35	65.00±1.58	71.10±4.32	51.43±3.31	59.96±5.13	50.00±0.17	–	–	–	–
ProtoNet	74.98±0.32	65.58±1.72	64.54±0.89	57.50±2.34	65.88±4.11	58.31±3.18	63.70±1.26	56.36±1.54	64.93±1.94	55.79±1.45
MAML	80.21±0.24	75.74±0.48	70.43±0.76	67.81±1.12	63.90±2.28	60.51±3.12	66.79±0.85	65.97±5.04	66.22±1.31	62.04±1.73
TPN	76.05±0.24	60.16±1.18	67.84±0.95	62.90±1.38	65.22±5.82	50.00±0.51	62.74±1.45	50.01±0.05	–	–
EGNN	81.21±0.16	79.44±0.22	72.87±0.73	70.79±0.95	65.20±2.08	62.18±1.76	63.65±1.57	61.02±1.94	69.92±1.85	62.14±1.58
IterRefLSTM	81.10±0.17	80.97±0.10	69.63±0.31	71.73±0.14	49.56±5.12	48.54±3.12	–	–	–	–
MHNfs	80.23±0.84	–	65.89±1.17	–	73.81±2.53	–	74.91±0.73	–	–	–
Pre-GNN	82.14±0.08	81.68±0.09	73.96±0.08	73.24±0.12	67.14±1.58	64.51±1.45	73.68±0.74	72.90±0.84	76.79±0.45	75.27±0.49
Meta-MGNN	82.94±0.10	82.13±0.13	75.43±0.21	73.36±0.32	68.99±1.84	65.54±2.13	76.27±0.57	72.43±0.85	72.58±0.34	72.51±0.35
Pre-PAR	84.93±0.11	83.01±0.09	78.08±0.16	74.46±0.29	69.96±1.37	66.94±1.12	75.12±0.84	73.63±1.00	73.71±0.61	72.49±0.61
Pre-GS-Meta	86.91±0.41	86.46±0.55	85.08±0.54	84.45±0.26	70.18±1.25	67.15±2.04	83.81±0.16	81.57±0.18	79.40±0.43	78.16±0.47
Pre-ADKF-IFT	86.06±0.35	80.97±0.48	70.95±0.60	62.16±1.03	95.74±0.37	67.25±3.87	76.22±0.13	71.13±1.15	80.21±0.27	76.62±0.73
Pre-PACIA	86.40±0.27	84.35±0.14	83.97±0.22	80.70±0.28	73.43±1.96	69.26±2.35	76.22±0.73	75.09±0.95	75.36±0.30	70.16±1.33
Pre-KRGTS	87.62±0.29	87.54±0.11	85.09±0.31	84.61±0.16	74.47±0.82	68.69±0.60	84.02±0.10	82.39±0.29	81.59±0.30	81.18±0.17
Pin-Tuning	91.56±2.57	85.71±1.33	93.41±3.52	84.36±1.73	73.33±2.00	66.44±2.50	84.94±1.09	80.08±1.21	81.26±0.46	71.98±1.55
Pre-UniMatch	86.35±0.13	–	80.34±0.45	–	86.35±0.76	–	81.63±0.73	–	–	–
MLTI	78.95±0.17	65.49±2.45	81.57±1.72	65.85±1.48	–	–	–	–	–	–
SMILE	76.23±0.69	60.69±1.24	71.80±0.79	56.49±1.49	–	–	–	–	–	–
ReCoG	93.52±1.47	91.01±2.03	93.98±1.53	90.06±1.17	95.94±1.37	81.96±0.42	88.85±1.39	86.05±1.74	82.57±1.42	82.39±1.64
CalibHyper	**96.82±1.70**	**95.96±2.47**	**96.89±2.33**	**94.93±1.08**	**97.19±1.21**	**92.93±1.51**	**90.01±0.93**	**89.64±1.37**	**90.37±2.33**	**90.15±1.57**

Baselines and reporting. Table 1 includes methods trained from scratch: Siamese (Koch et al., 2015), ProtoNet (Snell et al., 2017), MAML (Finn et al., 2017), TPN (Liu et al., 2019), EGNN (Kim et al., 2019), IterRefLSTM (Altae-Tran et al., 2017), and MHNfs (Schimunek et al., 2023). Methods with a pretrained encoder are Pre-GNN (Hu et al., 2020), Meta-MGNN (Guo et al., 2021), Pre-PAR (Wang et al., 2021), Pre-GS-Meta (Zhuang et al., 2023), Pre-ADKF-IFT (Chen et al., 2023), Pre-PACIA (Wu et al., 2024), Pre-KRGTS (Wang et al., 2025b), Pin-Tuning (Wang et al., 2024), and Pre-UniMatch (Li et al., 2025). MLTI (Yao et al., 2022) and SMILE (Liu et al., 2025b) are augmentation methods, and ReCoG (Wang et al., 2026) is the closest method. All baseline values and error terms are taken from Wang et al. (2026). We report ROC-AUC (%) averaged over the held-out properties of each dataset. Tables 1 and 2 give the mean and sample standard deviation over three runs with a fixed configuration per entry, and the ablations in Section 6.4 are single runs.

6.2 Main Results

CalibHyper exceeds the strongest reported baseline in all 26 dataset and shot conditions (Table 1; Table C.5 gives the nine ToxCast subsets). Two 10-shot ToxCast entries, CLD and Tanguay, come from an ablation setting with binary relation targets and the MSRHA gate fixed at $\gamma = 0$.

The margins follow the label statistics of Section 4. On PCBA, where joint observations are abundant and $\rho_{pq} \approx 1$, the margin over ReCoG is about eight points in both settings. NVS, whose median pair has only 22 joint observations, is the lowest-scoring subset for both methods and has among the smallest margins. Among the subsets, CEETOX has the second-largest margin in both settings. Its paired increase and decrease assays are strongly negatively dependent (Figure 2d); the marginal term $2c_{pq} - 1$ partly offsets this dependence in a binary score (Equation 3), whereas the calibrated score keeps the negative sign and MSRHA routes such relations through the negative channel. On Tox21, SIDER, MUV, and eight of the nine ToxCast subsets, the margin is larger in the 1-shot setting, where prediction relies more on the auxiliary context.

6.3 Component Ablation

All ablations in Table 2 use the same three-seed protocol, and the full model has the highest mean on every dataset. Removing MSRHA costs more than removing C4RC on all three datasets, most on MUV, yet neither module helps reliably on its own. With the four-state loss but no adapter, the score falls below that of removing both modules on all three datasets, so a relation loss that only shapes shared representations does not improve prediction. With the adapter but binary targets, the score exceeds that of removing both modules on SIDER and MUV but not on Tox21, where removing both scores above either single removal. The two modules therefore interact rather than add.

Table 2: Component ablation (10-shot ROC-AUC, %, three runs).

Ablation	Tox21	SIDER	MUV
CalibHyper (full)	**96.82±1.70**	**96.89±2.33**	**97.19±1.21**
w/o C4RC	90.08±2.96	90.78±3.33	93.82±2.19
w/o MSRHA	87.48±3.52	86.46±1.92	67.23±1.94
w/o C4RC and MSRHA	93.97±2.80	89.20±3.90	87.15±3.53
w/o chance correction	93.00±3.13	88.39±1.67	87.47±3.97
w/o sign separation	95.81±1.12	88.75±1.93	67.43±2.52

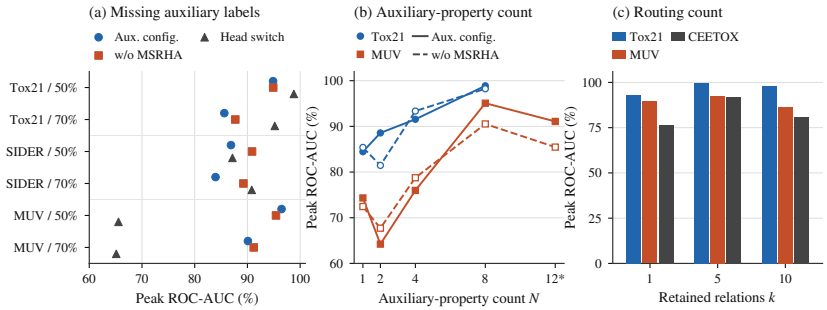


Figure 4: Single-run ablations, peak 10-shot ROC-AUC; details in Appendix C. (a) Masked auxiliary labels. (b) Auxiliary-property count N (12^*: 11 during MUV training). (c) Routing count k.

Inside the modules, keeping four states without chance correction improves on binary targets only on Tox21, so the four-state encoding alone does not explain the gains. Merging the signed channels gives the best ablated score on Tox21 but drops MUV to the level of the ablation without MSRHA. On MUV, rare positives leave almost no room for negative dependence, so negative scores carry little signal, and a separate channel keeps them apart from positive messages. The unsigned ablation also changes the normalization.

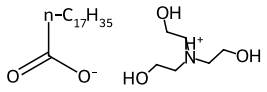
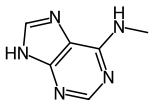
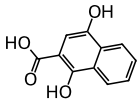
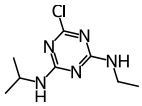
6.4 Ablations on Auxiliary Information and Routing

Missing auxiliary labels. On Tox21, SIDER, and MUV, we hide 50% or 70% of the observed auxiliary labels and ablate each dataset's base configuration ("Aux. config.") by removing MSRHA or switching its relation head between binary and four-state targets (Figure 4a). Single runs are compared within each dataset, not with the three-run means in Table 1. Masking removes joint observations, which lowers ρ_{pq} and makes the calibrated residuals noisier. On Tox21 the binary head gives the highest peak at both rates, while the base configuration and the ablation without MSRHA are nearly tied at 50%. On MUV, switching to four-state targets drops the peak to Pre-GNN's level at both rates. When positives are rare, the marginal term $2c_{pq} - 1$ in Equation 3 is close to 1, so binary scores send nearly every auxiliary property through the positive channel with similar weights, whereas the calibrated score is close to zero and routing rests on small residuals. On SIDER, the base configuration freezes γ at zero and thus matches the ablation without MSRHA up to the random state, yet their peaks differ by four to five points, the scale of run-to-run variation.

Auxiliary-property count. We vary N from 1 to 8 on Tox21 and from 1 to 12 on MUV (Figure 4b). More auxiliary properties generally raise the peak, although both MUV curves dip at $N = 2$ and decline at $N = 12$. With four or fewer auxiliary properties neither setting is consistently better, whereas in the three settings with $N \geq 8$ MSRHA gives the higher peak, so the adapter pays off once there are enough relations to route. The advantage holds only at the peak; by the end of training, these runs with MSRHA lose more than those without it.

Routing count. The routing count k sets how many relations pass messages to each molecule. Among $k = 1$, 5, and 10, Tox21, MUV, and CEETOX all peak at $k = 5$ (Figure 4c). A single

Table 3: Molecular cases (10-shot): positive-class score and the label predicted at threshold 0.5 (bold: correct; GT: ground truth). CEETOX target names omit the prefix CEETOX_H295R_.

Case	Molecule	Dataset and target	GT	ReCoG	Pin-Tuning	CalibHyper
C1		Tox21 SR-HSE	Neg.	0.99 Pos.	0.59 Pos.	7.39×10^{-6} **Neg.**
C2		Tox21 SR-HSE	Neg.	1.00 Pos.	0.57 Pos.	3.77×10^{-8} **Neg.**
C3		CEETOX ESTRONE_up	Neg.	1.00 Pos.	0.83 Pos.	7.98×10^{-5} **Neg.**
C4		CEETOX ESTRADIOL_up	Neg.	0.98 Pos.	0.87 Pos.	0.02 **Neg.**

relation discards complementary evidence, which costs most on CEETOX, while $k = 10$ admits weakly related properties and, on Tox21, removes selection altogether.

6.5 Molecular Cases

In the 10-shot setting on Tox21 and on the CEETOX subset of ToxCast, ReCoG (Wang et al., 2026), Pin-Tuning (Wang et al., 2024), and CalibHyper use the same support sets, query molecules, and auxiliary information, and query labels of the target are hidden before the forward pass. CalibHyper uses unsigned MSRHA on Tox21 and the full model on CEETOX, and the baselines use intermediate checkpoints with different training budgets. From 20,895 records of molecule and task pairs, Table 3 shows four negative molecules, selected post hoc, that CalibHyper classifies correctly and both baselines classify as positive.

For C3, the auxiliary labels DOC_up and PROG_up are positive while the target ESTRONE_up is negative. MSRHA routes both relations through the negative channel, so these positive auxiliary labels enter through the projection reserved for opposing relations, the case discussed for CEETOX in Section 6.2.

7 Conclusion

Label agreement mixes property dependence with class marginals and merges the two ordered disagreement states. CalibHyper replaces it with chance-corrected four-state residuals, shrunk when joint observations are few, and uses the predicted residuals to select, sign, and weight hyperedge messages for each molecule. On 26 conditions, CalibHyper and its ablation settings are competitive with the strongest reported baselines, with larger margins in the 1-shot setting. The margins follow the label statistics: large on PCBA, where joint observations are abundant, and on CEETOX, whose paired assays are negatively dependent, and small on NVS, where joint observations are scarce. The ablations show that the relation loss alone does not improve prediction and that the two modules interact, and routing a few relations per molecule works better than routing one or many.

Limitations. Labels missing not at random (Rubin, 1976; Little & Rubin, 2019) and extreme class marginals can bias the relation estimates. On MUV, rare positives and weak dependence leave the calibrated scores near zero, and four-state targets degrade under label masking. Gains vary by dataset and configuration, performance can decline late in training, and the main results cover three seeds.

AI USE STATEMENT

We used large language models to assist with writing experiment scripts and orchestrating runs, checking logs, aggregating numerical results, and revising and translating the text of the paper. The authors take responsibility for the methods and experimental conclusions, the accuracy of the content, and the final manuscript.

ETHICS STATEMENT

This work uses publicly available molecular property datasets and involves no human subjects or personal information. The intended applications are drug and materials screening.

REPRODUCIBILITY STATEMENT

Formal definitions and proofs are given in Appendices E, F, and G, and the experimental setup and full results are given in Appendix C. Appendix D describes the implementation, the ablation settings, and the computing environment. Code, processed data, and the command-line switches of all ablations are provided in the supplementary material.

REFERENCES

Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4):283–293, 2017.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Garrison Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. In *Uncertainty in Artificial Intelligence*, pp. 1352–1361, 2021. URL https://proceedings.mlr.press/v161/bachlechner21a.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5):423–429, 1993.

Haibo Chen, Xin Wang, Zeyang Zhang, Haoyang Li, Ling Feng, and Wenwu Zhu. AutoGFM: Automated graph foundation model with adaptive architecture customization. In *International Conference on Machine Learning*, pp. 9295–9315, 2025. URL https://proceedings.mlr.press/v267/chen25bp.html.

Wenlin Chen, Austin Tripp, and José Miguel Hernández-Lobato. Meta-learning adaptive deep kernel Gaussian processes for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KXRSh0sdVTP.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999, 2016.

Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. Molecular representations in AI-driven drug discovery: A review and practical guide. *Journal of Cheminformatics*, 12:56, 2020. URL https://doi.org/10.1186/s13321-020-00460-5.

Tyler Derr, Yao Ma, and Jiliang Tang. Signed graph convolutional networks. In *IEEE International Conference on Data Mining*, 2018.

Tianyu Fan, Lirong Wu, Yufei Huang, Haitao Lin, Cheng Tan, Zhangyang Gao, and Stan Z. Li. Decoupling weighing and selecting for integrating multiple graph pre-training tasks. In *International Conference on Learning Representations*, 2024.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI Conference on Artificial Intelligence*, pp. 3558–3565, 2019.

Yifan Feng, Shiquan Liu, Shihui Ying, Shaoyi Du, Zongze Wu, and Yue Gao. Self-supervised hypergraph training framework via structure-aware learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10160–10176, 2025.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. In-context learning for few-shot molecular property prediction. *arXiv preprint arXiv:2310.08863*, 2023. URL https://arxiv.org/abs/2310.08863.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135, 2017.

Yue Gao, Yifan Feng, Shiquan Liu, Xiangmin Han, Shaoyi Du, Zongze Wu, and Han Hu. Hypergraph foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (4):4063–4080, 2026.

Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V. Chawla. Few-shot graph learning for molecular property prediction. In *Proceedings of the Web Conference*, pp. 2559–2567, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Van Thuy Hoang and O-Joun Lee. Context-aware graph causality inference for few-shot molecular property prediction. *arXiv preprint arXiv:2601.11135*, 2026. URL https://arxiv.org/abs/2601.11135.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020.

Tinglin Huang, Ziniu Hu, and Rex Ying. Learning to group auxiliary datasets for molecule. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8e2571d13f432b301d4c5e3cc70227a6-Abstract-Conference.html.

Willard James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379, 1961.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D. Yoo. Edge-labeling graph neural network for few-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2019.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

Namkyeong Lee, Dongmin Hyun, Gyoung S. Na, Sungwon Kim, Junseok Lee, and Chanyoung Park. Conditional graph information bottleneck for molecular relational learning. In *International Conference on Machine Learning*, pp. 18852–18871, 2023. URL https://proceedings.mlr.press/v202/lee23e.html.

Ruifeng Li, Mingqian Li, Wei Liu, Yuhua Zhou, Xiangxin Zhou, Yuan Yao, Qiang Zhang, and Hongyang Chen. UniMatch: Universal matching from atom to task for few-shot drug discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/87eaaa8605a1a472d9a9756e7500517b-Abstract-Conference.html.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, third edition, 2019.

Gang Liu, Srijit Seal, John Arevalo, Zhenwen Liang, Anne E. Carpenter, Meng Jiang, and Shantanu Singh. Learning molecular representation in a cell. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/18610cbcd1da57854aa05ebdc5cd3168-Abstract-Conference.html.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One For All: Towards training one graph model for all classification tasks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/57faf5642eb06e0602b95f6aa989b38a-Abstract-Conference.html.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. In *International Conference on Learning Representations*, 2019.

Yonghao Liu, Mengyu Li, Fausto Giunchiglia, Lan Huang, Ximing Li, Xiaoyue Feng, and Renchu Guan. Dual-level mixup for graph few-shot learning with fewer tasks. In *Proceedings of the ACM on Web Conference*, pp. 2646–2656, 2025b.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, 2018. URL https://doi.org/10.1039/C8SC00148K.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 157–163, 1956.

Sebastian G. Rohrer and Knut Baumann. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of Chemical Information and Modeling*, 49(2):169–184, 2009.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Johannes Schimunek, Philipp Seidl, Lukas Friedrich, Daniel Kuhn, Friedrich Rippmann, Sepp Hochreiter, and Günter Klambauer. Context-enriched molecule representations improve few-shot drug discovery. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XrMWUuEevr.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 2018.

Sangwoo Seo, Sungwon Kim, and Chanyoung Park. Interpretable prototype-based graph information bottleneck. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Megan Stanley, John F. Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-Mol: A few-shot learning dataset of molecules. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Liang Wang, Qiang Liu, Shaozhen Liu, Xin Sun, Shu Wu, and Liang Wang. Pin-Tuning: Parameter-efficient in-context tuning for few-shot molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 70765–70792, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/82a9fb94035dad3ec007de4ad13c6748-Abstract-Conference.html.

Yaqing Wang, Abulikemu Abuduweili, Quanming Yao, and Dejing Dou. Property-aware relation networks for few-shot molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17441–17454, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/91bc333f6967019ac47b49ca0f2fa757-Abstract.html.

Zehong Wang, Zheyuan Zhang, Tianyi Ma, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Towards graph foundation models: Learning generalities across graphs via task-trees. In *International Conference on Machine Learning*, pp. 65518–65555, 2025a. URL https://proceedings.mlr.press/v267/wang25eq.html.

Zeyu Wang, Tianyi Jiang, Yao Lu, Xiaoze Bao, Shanqing Yu, Bin Wei, Qi Xuan, and Hong Wang. Knowledge-enhanced relation graph and task sampling for few-shot molecular property prediction. *Information Sciences*, 718:122357, 2025b. URL https://doi.org/10.1016/j.ins.2025.122357.

Zeyu Wang, Tianyi Jiang, Huanchang Ma, Yao Lu, Xiaoze Bao, Shanqing Yu, Qi Xuan, Shirui Pan, and Xin Zheng. Few-shot molecular property prediction: A survey. *arXiv preprint arXiv:2510.08900*, 2025c. URL https://arxiv.org/abs/2510.08900.

Zeyu Wang, Xin Zheng, Yao Lu, Shanqing Yu, Qi Xuan, and Shirui Pan. ReCoG: Relational and compact context graph learning for few-shot molecular property prediction. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61905.

Matthijs J. Warrens. On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73(4):777–789, 2008.

Shiguang Wu, Yaqing Wang, and Quanming Yao. PACIA: Parameter-efficient adapter for few-shot molecular property prediction. In *International Joint Conference on Artificial Intelligence*, pp. 5208–5216, 2024. URL https://www.ijcai.org/proceedings/2024/0576.pdf.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. URL https://doi.org/10.1039/C7SC02664A.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Huaxiu Yao, Linjun Zhang, and Chelsea Finn. Meta-learning with fewer tasks through task interpolation. In *International Conference on Learning Representations*, 2022.

Li Zhang, Dan Xu, Anurag Arnab, and Philip H. S. Torr. Dynamic graph message passing networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3726–3735, 2020.

Shuai Zhang, Junfeng Fang, Xuqiang Li, Hongxin Xiang, Jun Xia, Ye Wei, Wenjie Du, and Yang Wang. Iterative substructure extraction for molecular relational learning with interactive graph information bottleneck. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9f989633ffbd47a83caddacad0f0261f-Abstract-Conference.html.

Yifan Zhang, Shaoyi Du, Yifan Feng, Shihui Ying, and Yue Gao. Reinterpreting hypergraph kernels: Insights through homomorphism analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):466–481, 2026.

Zhiqiang Zhong and Davide Mottin. Automatic auxiliary task selection and adaptive weighting boost molecular property prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/61c2975281d60d3b1ce4cefc157d99df-Abstract-Conference.html.

Xiang Zhuang, Qiang Zhang, Bin Wu, Keyan Ding, Yin Fang, and Huajun Chen. Graph sampling-based meta-learning for molecular property prediction. In *International Joint Conference on Artificial Intelligence*, pp. 4729–4737, 2023. URL https://www.ijcai.org/proceedings/2023/0526.pdf.

List of Appendices

A Extended Related Work **16**

A.1 Few-Shot Molecular Property Prediction and Context-Aware Methods 16

A.2 Selecting, Weighting, and Compressing Auxiliary Information 16

A.3 Agreement Correction and Relational Structure 16

A.4 Evaluation Protocol and Scope of Comparison 16

B Notation **17**

C Experimental Setup and Additional Results **18**

C.1 Evaluation Protocol and Baseline Sources 18

C.2 Missing Auxiliary Labels and Auxiliary-Property Count 18

C.3 Routing Count 20

C.4 Results on ToxCast Subsets 20

D Implementation Details and Computing Environment **21**

E Formal Definitions and Assumptions **21**

E.1 Label Matrix and Empirical Law of a Property Pair 21

E.2 Assumptions 22

F Proofs **23**

F.1 Algebraic Decomposition of the Joint State 23

F.2 Proof of Proposition 1 23

F.3 Degeneracy of the Pair Term 23

F.4 Expectation of the C4RC Target 24

F.5 Swap Equivariance of the Relation Head 24

F.6 Identity at Initialization 24

G Further Analysis and Scope **25**

G.1 Marginal Component of the Binary Target 25

G.2 Finite Observations, Shrinkage, and Smoothing 25

G.3 Expressiveness of the Hyperedge Encoder 26

G.4 Label Use and Mean Signed Scores 26

H Label Statistics Behind the Relation Model **26**

H.1 Chance Agreement under Sparse Labels: MUV-713 and MUV-733 27

H.2 Direction of Disagreement: Increase and Decrease of the Same Readout 27

H.3 Molecule-Dependent Association: NR-AR and SR-p53 in Tox21 27

H.4 Support from Joint Observations and the Effect of a Single Label 27

A Extended Related Work

A.1 Few-Shot Molecular Property Prediction and Context-Aware Methods

David et al. (2020) review molecular representations. Metric-based methods predict through similarities or prototypes (Koch et al., 2015; Snell et al., 2017; Altae-Tran et al., 2017), and gradient-based meta-learning optimizes parameters that adapt well within a task (Finn et al., 2017; Guo et al., 2021). Molecular pretraining methods include Pre-GNN (Hu et al., 2020) and GROVER (Rong et al., 2020), and FS-Mol (Stanley et al., 2021) provides a separate benchmark for few-shot evaluation. Multitask drug discovery shares statistical strength across the property labels of the same molecule (Ramsundar et al., 2015; Mayr et al., 2018), and context-aware methods carry this idea over to unseen properties by using such auxiliary observations as context.

KRGTS (Wang et al., 2025b) enriches the relation graph with scaffold and functional-group similarity. CaMol (Hoang & Lee, 2026) uses a context graph of functional groups, molecules, and properties to guide atom masking and substructure extraction, and constructs distributional interventions under its assumed causal model. CalibHyper instead estimates how far jointly observed labels deviate from the independence baseline implied by their marginals.

A.2 Selecting, Weighting, and Compressing Auxiliary Information

Besides MolGroup and AutAuT, WAS (Fan et al., 2024) decouples the selection and the weighting of pretrained teachers at the instance level and then trains a student by distillation. MolGroup learns a routing mechanism through bilevel optimization, and AutAuT discovers auxiliary tasks with a large language model; the three methods select auxiliary datasets, auxiliary tasks, and pretrained teachers, respectively. CalibHyper works within the auxiliary property set of a given episode and learns relation residuals that generate molecule-conditioned messages.

Graph information bottleneck methods follow the information bottleneck principle (Tishby et al., 1999) and differ in the conditioning variable and in what they compress. Besides CGIB, IGIB-ISE, and InfoAlign, PGIB (Seo et al., 2023) combines the compression of key subgraphs with prototype similarity for interpretable prediction. These methods compress substructures of the input graph or molecular representations. Target-conditioned compression of auxiliary property nodes has also been studied in ReCoG (Wang et al., 2026). CalibHyper uses this form of context compression together with calibrated relation supervision and signed propagation.

A.3 Agreement Correction and Relational Structure

Warrens (2008) studies which association coefficients for 2×2 tables do not depend on the marginal distributions. CalibHyper applies chance correction to the four ordered label states and shrinks the result according to the number of joint observations.

Hypergraph learning can also be approximated with graph structures (Yadati et al., 2019), and relational graph convolutions use a separate transformation for each relation type (Schlichtkrull et al., 2018). Equivariant networks map each transformation of the input to a fixed transformation of the output (Cohen & Welling, 2016). Residual learning adds an increment to an existing representation (He et al., 2016), and a zero-initialized residual gate lets the new path start from the identity map (Bachlechner et al., 2021). Our relation head and residual gate follow these two designs (Propositions 4 and 5).

A.4 Evaluation Protocol and Scope of Comparison

Reproducible comparisons require explicit task splits, model configurations, sources of randomness, and aggregation rules (Bouthillier et al., 2021; Pineau et al., 2021). Our main comparison and component analysis use mean ROC-AUC and sample standard deviation across independent runs; Appendix C.1 specifies the reporting protocol and distinguishes it from the single-run ablations.

Graph prompting methods and graph foundation models target adaptation across graphs or tasks. OFA (Liu et al., 2024) unifies feature spaces through text attributes and organizes classification and few-shot tasks at different levels with prompt graphs, GIT (Wang et al., 2025a) unifies pretraining

and in-context learning with task trees, and AutoGFM (Chen et al., 2025) chooses the architecture according to the features of the downstream graph. We evaluate 1-shot and 10-shot ROC-AUC on unseen properties.

B NOTATION

Table B.1 lists the notation of the main text, and Table B.2 the additional symbols of Appendices E to G.

Table B.1: Notation of the main text.

Symbol	Meaning
$\mathcal{M}, \mathcal{T}$	set of molecules, set of properties
$\mathcal{T}_{\text{train}}, \mathcal{T}_{\text{test}}$	disjoint meta-training and meta-test properties
$y_{i,p} \in \{0, 1, \star\}$	label of molecule i for property p; $\star$ means unmeasured
$o_{i,p}$	observation indicator $\mathbf{1}[y_{i,p} \neq \star]$
$E_\tau = (S_\tau, Q_\tau)$	episode with target τ, its support set and query set
A_τ, N	auxiliary properties of the episode and their number
K, M	support molecules per class, query molecules
G_τ, V_τ^m, V_τ^p	context graph, its molecule nodes and property nodes
F_{mol}, d	pretrained molecule encoder, embedding width ($d = 300$)
F_{rel}, F_{pred}	relation head, property predictor
$H = \{h_v\}$, $H' = \{h'_v\}$	output of the context encoder, representations read by the predictor
λ_a	bottleneck gate of auxiliary property a
$e(u, v) \in \{0, 1\}^4$	one-hot joint label state in the order $(00, 01, 10, 11)$
n_{pq}	number of molecules observed for both p and q
$\hat{\pi}_{p\|q}$, α	smoothed positive rate of p on jointly observed molecules, Laplace constant
b_{pq}	four-state independence baseline
ρ_{pq}, n_0	reliability of joint observations, shrinkage constant
$r_{i,p,q} \in \mathbb{R}^4$	chance-corrected four-state relation target, Equation 2
$\hat{r}_{i,p,q}$	prediction of the swap-equivariant head
ξ_{ipq}, δ_{ipq}	swap-invariant features, antisymmetric scalar splitting the disagreement states
$d_{i\tau a}$	signed agreement, Equation 5
$\beta_{i\tau a}$, k	routing score, auxiliary properties kept per molecule and target
$w_e^{\pm}$, z_e	signed hyperedge weights, hyperedge encoding
$\text{rl}(v)$	role of a node (molecule or target), which selects the write-back projection
γ	residual gate of the adapter, initialized at zero
η_{rel}	weight of the relation loss
κ_{pq}, a_{pq}, c_{pq}	Cohen's κ, observed agreement, chance agreement

Table B.2: Additional symbols of Appendices E to G.

Symbol	Meaning
I_{pq}, P_{pq}	joint observation set and the empirical law of the pair on it (Definition E.2)
$\pi_{p\|q}$	unsmoothed positive rate of p on jointly observed molecules (Definition E.2)
b^*_{pq}, $r^*_{i,p,q}$	baseline and ideal target built from exact unsmoothed marginals (Definitions E.4 and E.6)
$\chi_\varnothing$, χ_p, χ_q, χ_{pq}	Walsh basis of $\mathbb{R}^4$, with $\chi_{pq} = \chi_p \odot \chi_q$ (Definition E.5)
s_p, s_q	labels in ± 1 form, $s_p = 1 - 2Y_p$ (Definition E.5)
P_{swap}, $\mathcal{S}$	swap acting on output coordinates and on inputs (Definition E.7)
$d(r)$, $\text{dir}(r)$	signed agreement and directional component (Definition E.8)
Δ_{pq}	agreement beyond chance, $a_{pq} - c_{pq}$ (Definition E.3)
ω^2, σ^2	prior variance of Δ_{pq}, scale of its sampling variance (Appendix G.2)
$\mathcal{P}$, $\Delta_h^{(3)}$	pairwise decomposable functions, mixed third-order difference (Definition E.10)
Π_0	zero-sum subspace of $\mathbb{R}^4$ (Appendix F.5)

C Experimental Setup and Additional Results

C.1 Evaluation Protocol and Baseline Sources

Task splits and sampling. Testing covers all held-out properties: 3 on Tox21, 6 on SIDER, 5 on MUV, and 10 on PCBA, and on the ToxCast subsets 10 on APR, 40 on ATG, 31 on BSK, 4 on CEETOX, 5 on CLD, 39 on NVS, 4 on OT, 20 on TOX21, and 5 on Tanguay. The support set usually contains K molecules per class. When the minority class has at most K molecules, one minority molecule is kept for the query set, and molecules of the other class fill the support set up to $2K$ molecules.

Metrics. On each dataset, ROC-AUC is averaged over the held-out properties. For ToxCast as a whole, each run averages the nine subsets equally at the same evaluation episode, and Table 1 reports its three-run mean; the entries of Table C.5 are separate three-run means. AP is computed with the `average_precision_score` function of scikit-learn. Reported differences are calculated before rounding; displayed experimental results and error terms use two decimal places.

Reporting. The CalibHyper results in Tables 1, 2, and C.5 are means over three independent runs. The model configuration is fixed within each entry; Table C.1 lists the best-performing hyperparameters of the main experiments. The complete model and ablations use the same evaluation protocol. For run scores x_1, x_2, x_3, the reported center is $\bar{x} = \frac{1}{3}\sum_{s=1}^{3} x_s$ and the $\pm$ term is the sample standard deviation $\sqrt{\frac{1}{2}\sum_{s=1}^{3}(x_s - \bar{x})^2}$. Baseline numbers retain their original definitions: Pin-Tuning, GS-Meta, and PACIA report means and standard deviations over ten seeds, whereas ReCoG does not specify its repetition count or the definition of its error term. The ablations of Section 6.4 are single runs with 2,000 training episodes and evaluation every 100 episodes, using deterministic evaluation and single-process data loading; they are compared within each group, not with the three-run means.

Archived configuration and checkpoint selection. The archived searches refined dataset- and shot-specific configurations over successive rounds, varying learning rates, loss weights, smoothing, shrinkage, and adapter settings across CalibHyper and its ablations. Each recorded 2,000-episode run was evaluated at 100-episode intervals. The checkpoint with the highest mean ROC-AUC across meta-test properties was selected from these 20 evaluations, and configurations were ranked by this score. No separate validation split was used for selection. The archived three-seed summaries average the scores at the three independently selected checkpoints and report their sample standard deviation. For the archived ToxCast aggregate, the nine subsets are averaged at each common evaluation episode, and the episode with the highest aggregate score is selected. These statistics retain the effect of test-based selection.

Scope of the selection records. The archived runs and the released training loop document this selection procedure, but have not been matched to the individual runs behind the current main-table summaries. Their configuration and checkpoint choices therefore remain unverified by these records.

Baseline sources. The nineteen compared methods in Table 1 come from Wang et al. (2026): sixteen baselines and ReCoG from its Table 1, and MLTI and SMILE from its Table 11, which reports only 1-shot and 10-shot ROC-AUC on Tox21 and SIDER. Values, $\pm$ terms, and missing entries follow the original. The eight ToxCast baselines in Table C.5 come from its Tables 8 and 9.

Table C.1: Hyperparameters of the main experiments; the rest follow the anonymous repository.

GIN layers / width d	5 / 300	Training episodes	2,000	Laplace constant α	1
Query batch (train/test)	16 / 64	Routing count k	5	Shrinkage constant n_0	5

C.2 Missing Auxiliary Labels and Auxiliary-Property Count

Tables C.2 and C.3 list the single-run ablations on missing auxiliary labels and on the number of auxiliary properties, 18 runs each (Figure 4a, b). Each dataset has its own ablation base ("Aux.

config."): C4RC with unsigned MSRHA on Tox21, binary relations with MSRHA on MUV, and binary relations with γ fixed at zero on SIDER. The ablations keep the parameters and seeds of each group, and switching the head changes between binary and four-state targets. Peak is the highest value over all evaluations, Last-5 is the mean ROC-AUC of the last five, and Final is the last value.

Table C.2: Missing auxiliary labels in the 10-shot setting (%).

		ROC-AUC			AP	
Masked	Ablation setting	Peak	Last-5	Final	Peak	Final
Tox21						
50%	C4RC, unsigned MSRHA	94.87	73.73	86.32	85.53	40.31
50%	w/o MSRHA	94.89	79.70	90.07	66.13	49.88
50%	Head switched to binary	98.80	88.44	80.66	93.30	30.17
70%	C4RC, unsigned MSRHA	85.62	72.15	57.30	41.84	22.39
70%	w/o MSRHA	87.70	75.99	87.70	44.59	44.59
70%	Head switched to binary	95.19	79.22	95.19	77.19	66.97
SIDER						
50%	Binary, γ fixed at 0	86.88	78.16	73.01	83.68	70.07
50%	w/o MSRHA	90.88	78.47	72.40	88.23	67.58
50%	Head switched to C4RC	87.16	82.62	87.16	79.75	77.74
70%	Binary, γ fixed at 0	83.95	68.58	40.80	80.96	57.07
70%	w/o MSRHA	89.24	82.57	89.24	85.12	85.12
70%	Head switched to C4RC	90.82	85.11	87.56	83.53	77.80
MUV						
50%	Binary, MSRHA	96.48	83.20	65.00	43.05	23.75
50%	w/o MSRHA	95.42	72.92	59.26	50.93	16.62
50%	Head switched to C4RC	65.54	60.06	58.09	1.23	0.30
70%	Binary, MSRHA	90.10	79.79	61.91	34.71	21.75
70%	w/o MSRHA	91.19	87.97	87.02	23.22	23.22
70%	Head switched to C4RC	65.12	62.11	60.10	1.51	0.34

Table C.3: Number of auxiliary properties in the 10-shot setting (%).

		ROC-AUC			AP	
N	Ablation setting	Peak	Last-5	Final	Peak	Final
Tox21						
1	C4RC, unsigned MSRHA	84.46	82.58	82.68	38.22	34.26
1	w/o MSRHA	85.39	82.44	81.59	38.17	32.28
2	C4RC, unsigned MSRHA	88.57	83.13	82.84	47.35	36.14
2	w/o MSRHA	81.46	74.68	75.29	32.45	24.34
4	C4RC, unsigned MSRHA	91.55	79.00	87.60	54.61	54.52
4	w/o MSRHA	93.35	86.51	90.16	55.49	45.57
8	C4RC, unsigned MSRHA	98.88	67.67	37.79	90.43	7.05
8	w/o MSRHA	98.23	79.63	91.55	86.83	75.20
MUV						
1	Binary, MSRHA	74.33	64.33	61.49	0.88	0.37
1	w/o MSRHA	72.45	62.95	61.13	0.89	0.30
2	Binary, MSRHA	64.23	53.35	52.66	0.40	0.17
2	w/o MSRHA	67.72	56.60	59.50	0.47	0.23
4	Binary, MSRHA	75.98	71.26	73.75	8.03	0.90
4	w/o MSRHA	78.76	71.44	69.17	8.84	2.37
8	Binary, MSRHA	95.08	66.27	64.28	42.10	21.43
8	w/o MSRHA	90.53	85.89	82.42	28.12	23.30
12	Binary, MSRHA	91.08	54.13	44.21	38.51	8.69
12	w/o MSRHA	85.46	54.29	54.29	22.42	8.69

The count sweep uses seed 202. On Tox21, $N = 8$ uses eight auxiliary properties in both meta-training and testing; on MUV, $N = 12$ uses 11 in meta-training and 12 at test time.

Peak and final values often disagree. Many runs reach their peak and then degrade, in some cases below chance level. The drop is not specific to MSRHA: it also occurs without MSRHA, for example on MUV at 50% masking, and for the SIDER ablation base, whose gate is frozen at zero. In the count sweep with $N \geq 8$, however, the runs with MSRHA have the higher peak in all three settings but lose considerably more by the end of training; on Tox21 with $N = 8$, the final ROC-AUC of the ablation base falls below chance level while the ablation without MSRHA stays above 90. These peak and final metrics show that stable late training remains an open issue.

C.3 Routing Count

Table C.4 lists the single-run peak ROC-AUC of the routing ablation, with other settings fixed.

Table C.4: Peak ROC-AUC (%) for different routing counts k in the 10-shot setting.

k	Tox21	MUV	CEETOX
1	93.01	89.43	76.38
5	99.47	92.32	91.89
10	97.87	86.42	80.69

C.4 Results on ToxCast Subsets

Table C.5 uses the same experimental settings as Table 1. The best baseline is Pre-KRGTS on NVS and ReCoG elsewhere. Daggers mark the 10-shot CLD and Tanguay entries with inactive residual paths.

Table C.5: ROC-AUC (%) on the nine ToxCast subsets. CalibHyper: mean over three runs. Baselines are the means reported in Tables 8 and 9 of Wang et al. (2026); method groups follow Table 1. †Ablation setting with binary relation targets and the MSRHA gate fixed at $\gamma = 0$ (Proposition 5).

Method	APR	ATG	BSK	CEETOX	CLD	NVS	OT	TOX21	Tanguay
10-shot									
ProtoNet	73.58	59.26	70.15	66.12	78.12	65.85	64.90	68.26	73.61
MAML	72.66	62.09	66.42	64.08	74.57	66.56	64.07	68.04	77.12
EGNN	80.33	66.17	73.43	66.51	78.85	71.05	68.21	76.40	85.23
Pre-PAR	86.09	72.72	82.45	72.12	83.43	74.94	71.96	82.81	88.20
Pre-GS-Meta	90.15	82.54	88.21	74.19	86.34	76.29	74.47	90.63	91.47
Pre-KRGTS	90.31	80.12	87.92	76.63	86.97	77.52	75.11	89.83	91.73
Pin-Tuning	92.78	83.58	89.49	75.96	87.70	76.33	75.56	90.80	92.25
ReCoG	92.97	84.37	89.95	85.85	92.63	77.19	87.50	92.35	96.83
Best baseline	92.97	84.37	89.95	85.85	92.63	77.52	87.50	92.35	96.83
CalibHyper	**94.51**	**86.56**	**92.90**	**90.58**	**94.30**†	**79.32**	**92.71**	**94.21**	**98.96**†
1-shot									
ProtoNet	57.08	54.92	53.92	60.25	66.25	54.87	63.11	58.27	58.32
MAML	64.59	55.45	60.36	61.02	66.22	59.84	62.15	59.52	60.92
EGNN	67.06	57.28	60.82	60.10	71.53	56.56	66.08	63.32	74.80
Pre-PAR	84.69	70.38	79.89	66.57	77.83	72.51	70.41	80.33	86.64
Pre-GS-Meta	89.49	81.69	87.28	68.55	78.69	74.36	73.56	89.46	91.10
Pre-KRGTS	89.45	79.54	86.84	72.50	81.21	76.63	73.68	89.51	92.15
Pin-Tuning	87.08	80.67	83.12	71.91	79.15	71.29	73.17	84.27	90.08
ReCoG	94.39	84.92	89.23	76.71	87.79	76.29	77.89	91.79	95.40
Best baseline	94.39	84.92	89.23	76.71	87.79	76.63	77.89	91.79	95.40
CalibHyper	**95.76**	**95.70**	**93.63**	**87.61**	**97.80**	**79.20**	**89.32**	**99.03**	**99.88**

D Implementation Details and Computing Environment

Model. The relation head and the adapter are built after the base encoder and the predictor are initialized, and both use LayerNorm. With binary targets, the extra relation pass of MSRHA freezes the updates of BatchNorm running statistics but still uses batch statistics and dropout. The C4RC loss is skipped when an episode has no valid triple. Top-k routing is discontinuous where the selected set changes; for a fixed set, the weights of the retained edges are differentiable.

Complexity and overhead. Let d_r and d_h be the hidden widths of the relation head and the adapter. Counting dense linear layers, scoring candidate triples costs $O(|V_\tau^m|\, N\, d\, d_r)$, and encoding and writing back the retained edges costs $O(|V_\tau^m| \min(k, N)\, d\, d_h)$, plus the cost of top-$k$ selection; by default $d_r = d$. Relation supervision enumerates the valid property pairs on support molecules and the valid auxiliary pairs, excluding the target column, in each query minibatch, so there are at most $O(|V_\tau^m|N^2)$ triples. On Tox21, the backbone and prediction components have 3.02M parameters, while C4RC and MSRHA add 588K (19.50%), with 271K in the adapter and 317K in the four-state head.

Ablation settings. Besides the full model, the ablations replace C4RC with binary relation targets (w/o C4RC), remove MSRHA, remove chance correction and shrinkage while keeping four states (w/o chance correction), use unsigned MSRHA, which merges the two signed channels (w/o sign separation), or freeze the MSRHA gate at $\gamma = 0$ instead of learning it. The unsigned implementation sets the weight of the positive channel to $\lambda_a|d_{i\tau a}|$ and uses the normalizing mass $\lambda_a(|d_{i\tau a}| + \mathrm{ReLU}(-d_{i\tau a}))$, so edges with negative scores count twice in the denominator; this ablation therefore changes both the sign handling and the normalization. The implementation without chance correction regresses the one-hot state and keeps the zero-sum output. For a zero-sum prediction $\hat{r}$, $\|\hat{r} - e\|^2 = \|\hat{r} - (e - \frac{1}{4}\mathbf{1})\|^2 + \frac{1}{4}$, so it fits the centered one-hot vector up to a constant loss; it also removes the reliability shrinkage.

Gates and evaluation batches. Auxiliary gates are used only in bottleneck-augmented training and are set to 1 otherwise. The auxiliary runs use relation dropout of 0.1, so exact swap equivariance holds only at evaluation. Benchmark queries are batched in true-label class order; the cases in Section 6.5 use a fixed random order. Predictions depend on batch composition through the context graph and target-node aggregation.

Environment. The experiments used eight NVIDIA A800 80GB GPUs with PyTorch 2.7.1, PyTorch Geometric 2.6.1, and Python 3.10. Additional runs used eight Tesla V100S 32GB GPUs with Python 3.9.23 and ran two jobs per GPU.

E Formal Definitions and Assumptions

E.1 Label Matrix and Empirical Law of a Property Pair

Definition E.1 (Label matrix and observation mask)**.** The label matrix is $Y \in \{0, 1, \star\}^{|\mathcal{M}|\times|\mathcal{T}|}$, and the observation indicator is $o_{i,p} = \mathbf{1}[y_{i,p} \neq \star]$. The symbol $\star$ marks an unmeasured entry, and all sums run over observed entries only.

Definition E.2 (Joint observation set and empirical law of a pair)**.** For an ordered pair (p, q) with $p \neq q$, let

$$I_{pq} = \{\, i : o_{i,p} = o_{i,q} = 1 \,\}, \qquad n_{pq} = |I_{pq}|,$$

and assume $n_{pq} \geq 1$. Draw $i \sim \mathrm{Unif}(I_{pq})$ and treat $Y_p := y_{i,p} \in \{0, 1\}$ and $Y_q := y_{i,q} \in \{0, 1\}$ as random variables; their joint law is the empirical law P_{pq} of the pair. Unless stated otherwise, $\mathbb{E}$, Var, and Cov are taken under P_{pq}. The conditional marginals, the observed agreement, and the chance agreement are

$$\pi_{p|q} = \mathbb{E}[Y_p], \qquad \pi_{q|p} = \mathbb{E}[Y_q], \qquad a_{pq} = P_{pq}(Y_p = Y_q),$$

$$c_{pq} = (1 - \pi_{p|q})(1 - \pi_{q|p}) + \pi_{p|q}\, \pi_{q|p}.$$

Proposition 1 abbreviates $\pi_{p|q}$ and $\pi_{q|p}$ as π_p and π_q.

Definition E.3 (Cohen's κ and agreement beyond chance)**.** If $c_{pq} < 1$,

$$\kappa_{pq} = \frac{a_{pq} - c_{pq}}{1 - c_{pq}}, \qquad \Delta_{pq} := a_{pq} - c_{pq} = (1 - c_{pq})\,\kappa_{pq}.$$

We call Δ_{pq} the agreement beyond chance.

Definition E.4 (Smoothed marginals, reliability, and independence baseline)**.** Let $n^{+}_{p|q} = |\{\, i \in I_{pq} : y_{i,p} = 1 \,\}| = n_{pq}\,\pi_{p|q}$. With a Laplace constant $\alpha > 0$ and a shrinkage constant $n_0 > 0$,

$$\hat{\pi}_{p|q} = \frac{n^{+}_{p|q} + \alpha}{n_{pq} + 2\alpha}, \qquad \rho_{pq} = \frac{n_{pq}}{n_{pq} + n_0} \in (0, 1),$$

$$b_{pq} = \Big[(1 - \hat{\pi}_{p|q})(1 - \hat{\pi}_{q|p}),\ (1 - \hat{\pi}_{p|q})\hat{\pi}_{q|p},\ \hat{\pi}_{p|q}(1 - \hat{\pi}_{q|p}),\ \hat{\pi}_{p|q}\hat{\pi}_{q|p}\Big] \in \Delta^3.$$

The baseline built from the unsmoothed marginals $\pi_{p|q}$ and $\pi_{q|p}$ is denoted b^{*}_{pq}; it is the product of the marginals of P_{pq}. Appendix G.2 bounds the smoothing error.

Definition E.5 (Four-state encoding and Walsh basis)**.** The vector $e(u, v) \in \{0, 1\}^4$ is the one-hot encoding of the joint state in the order $(00, 01, 10, 11)$, where the first digit is the label of p and the second that of q:

$$e(u, v) = \big[(1 - u)(1 - v),\ (1 - u)v,\ u(1 - v),\ uv\big].$$

On $\mathbb{R}^4$ we take the four vectors

$$\chi_{\varnothing} = (1, 1, 1, 1),\ \ \chi_p = (-1, -1, 1, 1),\ \ \chi_q = (-1, 1, -1, 1),\ \ \chi_{pq} = \chi_p \odot \chi_q = (1, -1, -1, 1).$$

They are pairwise orthogonal with $\|\chi\|_2^2 = 4$ and form the Walsh-Hadamard basis. Labels in ± 1 form are $s_p = 1 - 2Y_p$ and $s_q = 1 - 2Y_q$.

Definition E.6 (C4RC target)**.** For $i \in I_{pq}$,

$$r_{i,p,q} = \rho_{pq}\big[\, e(y_{i,p}, y_{i,q}) - b_{pq} \,\big] \in \mathbb{R}^4.$$

Replacing b_{pq} in Equation 2 with b^{*}_{pq} gives the ideal target $r^{*}_{i,p,q}$.

Definition E.7 (Swap actions, invariance, and equivariance)**.** Let $P_{\mathrm{swap}} \in \mathbb{R}^{4\times 4}$ be the permutation matrix that exchanges the second and third coordinates, that is, the states 01 and 10, and fixes the first and fourth; it is an involution, $P^2_{\mathrm{swap}} = I$. Let $\mathcal{S}(h_i, h_p, h_q) = (h_i, h_q, h_p)$ be the swap on inputs. A map $F : \mathbb{R}^{3d} \to \mathbb{R}^4$ is *swap-invariant* if $F \circ \mathcal{S} = F$ and *swap-equivariant* if $F \circ \mathcal{S} = P_{\mathrm{swap}} F$.

Definition E.8 (Signed agreement and directional component)**.** For any $r \in \mathbb{R}^4$,

$$d(r) = \langle r, \chi_{pq} \rangle = (r_{00} + r_{11}) - (r_{01} + r_{10}), \qquad \mathrm{dir}(r) = r_{01} - r_{10} = \big\langle r, \tfrac{1}{2}(\chi_q - \chi_p) \big\rangle.$$

The signed agreement d drives routing and the sign channels and is invariant under the swap; dir distinguishes the two disagreement states and changes sign under the swap.

Definition E.9 (Ordered ternary hyperedge and roles)**.** A hyperedge is an ordered triple $e = (m_i, \tau, a)$ whose entries play the roles of molecule, target, and auxiliary property. The hyperedge reads all three nodes and writes to $\mathrm{wr}(e) = \{m_i, \tau\}$.

Definition E.10 (Pairwise decomposable class and mixed third-order difference)**.** A function $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to \mathbb{R}$ is *pairwise decomposable* if there exist ψ_{12}, ψ_{13}, and ψ_{23} such that

$$f(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) + \psi_{13}(x_1, x_3) + \psi_{23}(x_2, x_3);$$

we denote this class by $\mathcal{P}$. For a step h, the forward difference is $(\Delta^h_j f)(x) = f(\dots, x_j + h, \dots) - f(x)$, and the mixed third-order difference is $\Delta^{(3)}_h = \Delta^h_1 \Delta^h_2 \Delta^h_3$.

E.2 Assumptions

(A1) Meaning of expectations. Appendices F and G take expectations under the empirical law P_{pq} by default. The sampling and shrinkage analysis in Appendix G.2 and the analysis of evaluation scores in Appendix G.4 use the randomness defined there.

(A2) Superpopulation and i.i.d. sampling. The sampling bound in Appendix G.2 assumes that jointly observed label pairs are independent and identically distributed, and the shrinkage analysis uses the prior and conditional noise model given there. Algebraic identities use the empirical law.

(A3) Exact marginals. Statements about a zero mean under independence use the exact unsmoothed baseline b^*_{pq}; Appendix G.2 bounds the bias caused by replacing it with b_{pq}.

(A4) Nondegeneracy. The analysis of κ_{pq} assumes $0 < \pi_{p|q}, \pi_{q|p} < 1$, $c_{pq} < 1$, and $n_{pq} \geq 1$. The case $c_{pq} = 1$ means that both properties are constant with the same value; Cohen's κ_{pq} is then undefined, while the C4RC target is still computed with the joint observation mask.

(A5) Inputs and randomness of the head. The relation head reads context representations $h_i, h_p, h_q \in \mathbb{R}^d$, which may carry support and auxiliary label information. Expressive limits are stated for fixed node representations, and exact equivariance assumes a deterministic forward pass or a shared random mask.

F Proofs

We use the definitions and assumptions of Appendix E.

F.1 Algebraic Decomposition of the Joint State

Lemma F.1 (Walsh decomposition). *With $s_u = 1 - 2u$ and $s_v = 1 - 2v$,*

$$e(u, v) = \tfrac{1}{4}\big[\chi_\varnothing - s_u\chi_p - s_v\chi_q + s_u s_v \chi_{pq}\big]. \tag{F.1}$$

Proof. By Definition E.5, the inner products of $e(u, v)$ with $\chi_\varnothing$, χ_p, χ_q, and χ_{pq} are 1, $-s_u$, $-s_v$, and $s_u s_v$. The basis is orthogonal and each vector has squared norm 4, which gives the expansion. Taking the expectation under a product measure with marginals (π_1, π_2) gives the corresponding coefficients of the independence baseline, whose second-order coefficient is $(1-2\pi_1)(1-2\pi_2)$. □

F.2 Proof of Proposition 1

Write $t = \mathbb{E}[Y_p Y_q]$, $\pi_1 = \pi_{p|q}$, and $\pi_2 = \pi_{q|p}$. Summing over the four states of two binary labels,

$$a_{pq} = 1 - \pi_1 - \pi_2 + 2t, \qquad c_{pq} = 1 - \pi_1 - \pi_2 + 2\pi_1\pi_2,$$

so that $a_{pq} - c_{pq} = 2(t - \pi_1\pi_2) = 2\,\mathrm{Cov}(Y_p, Y_q)$. Since $y_{\text{rel}} = \mathbf{1}[Y_p \neq Y_q]$, we have $\mathbb{E}[y_{\text{rel}}] = 1 - a_{pq} = (1-c_{pq})(1-\kappa_{pq})$. Lemma F.1 gives Equation 1, and $\mathrm{Cov}(1-2Y_p, 1-2Y_q) = 4\,\mathrm{Cov}(Y_p, Y_q)$ gives the remaining claims. □

F.3 Degeneracy of the Pair Term

Proposition 2 (Degeneracy of the pair term). *Let the input of the relation head be $u = [h_i \| h_p] \odot [h_i \| h_q]$, where $h_i, h_p, h_q \in \mathbb{R}^d$. Then*

$$u = \big[\, h_i \odot h_i \;\|\; h_p \odot h_q \,\big] \in \mathbb{R}^{2d}. \tag{F.2}$$

Hence (i) the first d coordinates do not depend on the property pair; (ii) the last d coordinates do not depend on the molecule; (iii) u is invariant under the swap $p \leftrightarrow q$; (iv) the pre-activation of the first linear layer of the head decomposes additively as $W^{(1)}(h_i \odot h_i) + W^{(2)}(h_p \odot h_q) + b$, with no explicit bilinear cross term between molecule and property; and (v) every swap-equivariant four-state output F computed from u satisfies $F_{01} = F_{10}$ and $\|F - r\|_2^2 \geq \frac{1}{2}(r_{01} - r_{10})^2$ for any ordered target r.

Later nonlinear layers can still combine molecule and property features, so the additive structure in (iv) concerns only the first pre-activation; the swap invariance in (iii) carries over to every function of u.

Proof. The Hadamard product acts blockwise on the concatenation, which gives Equation F.2. Writing the first-layer weight as $[W^{(1)}\ W^{(2)}]$ gives the additive decomposition of the pre-activation.

Since $h_p \odot h_q = h_q \odot h_p$, the input does not change when the two properties are swapped, so every deterministic function F computed from it satisfies $F \circ \mathcal{S} = F$. If the output is also required to be swap-equivariant, then $F = P_{\mathrm{swap}} F$, hence $F_{01} = F_{10}$, and for any ordered target r

$$\|F - r\|_2^2 \geq \tfrac{1}{2}(r_{01} - r_{10})^2.$$

To see this, write $\bar{F} = F_{01} = F_{10}$ and use $(\bar{F} - x)^2 + (\bar{F} - y)^2 \geq (x - y)^2/2$; the other error terms are nonnegative. The bound continues to hold after taking the expectation over the input distribution. □

F.4 Expectation of the C4RC Target

Proposition 3 (Unbiasedness under independence)**.** *If p and q are independent on the jointly observed molecules and the marginals are estimated exactly, then $\mathbb{E}[r_{i,p,q}] = \mathbf{0}$. More generally, with exact marginals and by Equation 1, the signed contrast of the target satisfies*

$$\mathbb{E}\big[(r_{00} + r_{11}) - (r_{01} + r_{10})\big] = \rho_{pq}\big(\mathbb{E}[s_p s_q] - \mathbb{E}[s_p]\,\mathbb{E}[s_q]\big) = 2\,\rho_{pq}\,(1 - c_{pq})\,\kappa_{pq}. \tag{F.3}$$

Proof. With exact joint-observation marginals, the four-state joint law differs from the product measure by $(t - \pi_1\pi_2)(1, -1, -1, 1)$, so

$$\mathbb{E}[r^*] = \rho_{pq}\,\mathrm{Cov}(Y_p, Y_q)\,\chi_{pq}.$$

Under independence the covariance vanishes, which gives unbiasedness. Taking the inner product with χ_{pq}, whose squared norm is 4, gives Equation F.3. The entries of the one-hot vector and of the baseline each sum to one, so the residual sums to zero at every point, also for the smoothed baseline.

By Lemma F.1, the inner products of the ideal residual with χ_p and χ_q are $2\rho_{pq}(Y_p - \pi_1)$ and $2\rho_{pq}(Y_q - \pi_2)$; they vary across molecules and have zero mean. The binary target $y_{\mathrm{rel}} = (1 - s_p s_q)/2$, in contrast, is an affine function of the second-order coefficient alone. □

Appendix G.2 gives the finite-sample shift of the smoothed estimate.

F.5 Swap Equivariance of the Relation Head

Proposition 4 (Exact swap equivariance of a deterministic head)**.** *For fixed deterministic maps f and g, every i, p, q, and every parameter value, $\hat{r}_{iqp} = P_{swap}\hat{r}_{ipq}$, where P_{swap} exchanges the coordinates* 01 *and* 10 *and fixes* 00 *and* 11*. In addition, $\sum_k \hat{r}_{ipq}[k] = 0$.*

Proof. With a deterministic forward pass or a shared mask, $\xi_{iqp} = \xi_{ipq}$ and $\delta_{iqp} = -\delta_{ipq}$, so Equation 4 only exchanges the coordinates 01 and 10. The centering matrix $I - \mathbf{1}\mathbf{1}^\top/4$ commutes with every permutation of coordinates, so it preserves equivariance and makes the output sum to zero. The target transforms in the same way: $e(y_q, y_p) = P_{\mathrm{swap}} e(y_p, y_q)$, $b_{qp} = P_{\mathrm{swap}} b_{pq}$, and $\rho_{qp} = \rho_{pq}$. □

The output assembly covers the zero-sum subspace $\Pi_0 = \{x \in \mathbb{R}^4 : \sum_k x_k = 0\}$: for any $x \in \Pi_0$, take $(\hat{r}_{00}, \hat{r}_{11}, u_c, \delta) = (x_{00}, x_{11}, x_{01} + x_{10}, x_{01} - x_{10})$. The symmetric features can collide, however. The property pairs $((1,3),(2,4))$ and $((1,4),(2,3))$ have the same sum, elementwise product, and absolute difference, and therefore the same symmetric features for every h_i. The two pairs are not related by a swap, yet the symmetric channel cannot tell them apart.

F.6 Identity at Initialization

Proposition 5 (Identity at initialization)**.** *Fix the existing parameters, node representations, and random state, and assume that the messages and their derivatives are finite. When $\gamma = 0$, $H' = H$, and the input and output of the predictor, as well as the gradient of the current prediction loss with respect to the existing parameters, coincide with those of the same model with the adapter disabled.*

Proof. Fix the existing parameters θ, the adapter parameters θ_{ad}, the node representations, and the random state. Let $\mathcal{L}$ be the prediction loss of the current forward pass, and assume that the messages and the derivatives used are finite. At $\gamma = 0$ we have $H' = H + \gamma\Delta H = H$ and

$$\left.\frac{\partial H'}{\partial \theta}\right|_{\gamma=0} = \frac{\partial H}{\partial \theta}, \qquad \left.\frac{\partial \mathcal{L}}{\partial \gamma}\right|_{\gamma=0} = \langle \nabla_{H'}\mathcal{L}, \Delta H\rangle, \qquad \left.\frac{\partial \mathcal{L}}{\partial \theta_{\mathrm{ad}}}\right|_{\gamma=0} = 0.$$

The predictor reads the adapter output only through H', so at a zero gate its input, its output, and the prediction loss equal those with the path disabled. With the derivatives above and the chain rule, the prediction gradients of the existing parameters also coincide. □

The identity requires normalization inside the message: if LayerNorm were applied to the residual sum, a zero gate would give $H' = \mathrm{LN}(H)$.

G Further Analysis and Scope

G.1 Marginal Component of the Binary Target

Risk of constant regression. For a fixed property pair, let $\mu = \mathbb{E}[y_{\text{rel}}] = (1 - c_{pq})(1 - \kappa_{pq})$. A constant prediction v has mean squared error $\mu(1-\mu) + (v-\mu)^2$. The prediction $v = 1 - c_{pq}$, which uses only the marginals, has excess risk $(1 - c_{pq})^2\kappa_{pq}^2$ over the best constant, and with fixed marginals the derivative of the target mean with respect to κ_{pq} is $-(1 - c_{pq})$.

Marginal constraints. Let $\pi_1 = \pi_{p|q}$ and $\pi_2 = \pi_{q|p}$. Nonnegativity of the four state probabilities gives $t = P(Y_p = 1, Y_q = 1) \in [\max(0, \pi_1 + \pi_2 - 1), \min(\pi_1, \pi_2)]$, and substituting into $\kappa_{pq} = 2(t - \pi_1\pi_2)/(\pi_1 + \pi_2 - 2\pi_1\pi_2)$ gives the attainable range. For nondegenerate equal marginals $\pi_1 = \pi_2 = \pi \le 1/2$, $\kappa_{pq} \in [-\pi/(1-\pi), 1]$, so rare positives leave little room for negative κ; unequal marginals can also restrict the upper end.

Mean binary score. Map a binary prediction $\hat{y}_{\text{rel}}$ to the signed score $1 - 2\hat{y}_{\text{rel}}$. The best constant prediction under MSE, or the population average of the best conditional-mean prediction under the same distribution, then gives

$$\bar{d}_{pq}^{\,\mathrm{bin}} = 1 - 2\,\mathbb{E}[y_{\text{rel}}] = (2c_{pq} - 1) + 2(1 - c_{pq})\,\kappa_{pq}.$$

The mean score contains the marginal term $2c_{pq} - 1$, which is absent from Equation F.3.

G.2 Finite Observations, Shrinkage, and Smoothing

Sampling error. Under the i.i.d. assumption (A2), the observed agreement is a mean of Bernoulli indicators. Hoeffding's inequality (Hoeffding, 1963) gives, with probability at least $1 - \zeta$,

$$|\hat{a}_{pq} - a_{pq}^{\mathrm{pop}}| \le \sqrt{\frac{\ln(2/\zeta)}{2n_{pq}}},$$

where a_{pq}^{pop} is the agreement rate of the jointly observed population.

Linear shrinkage. Suppose that the signal Δ across property pairs satisfies $\mathbb{E}\Delta = 0$ and $\mathrm{Var}(\Delta) = \omega^2 > 0$, and that its estimate satisfies $\mathbb{E}[\hat{\Delta} \mid \Delta] = \Delta$ and $\mathrm{Var}(\hat{\Delta} \mid \Delta) = \sigma^2/n$ with $n, \sigma^2 > 0$. For the shrunk estimate $\rho\hat{\Delta}$, expanding the square and using conditional unbiasedness gives

$$\mathrm{MSE}(\rho) = (1-\rho)^2\omega^2 + \rho^2\sigma^2/n, \qquad \rho^\star = \frac{n}{n + \sigma^2/\omega^2}.$$

The quadratic is strictly convex, and setting its derivative to zero gives the optimal shrinkage constant σ^2/ω^2. In the experiments, n_0 is treated as a hyperparameter.

Smoothing shift. Let $n = n_{pq}$. Subtracting the two estimates in Definition E.4 gives $\hat{\pi} - \pi = \alpha(1 - 2\pi)/(n + 2\alpha)$. Each coordinate of the baseline is a product of two factors in $[0, 1]$, so the triangle inequality for differences of products gives

$$\|b_{pq} - b^*_{pq}\|_\infty \leq \frac{2\alpha}{n + 2\alpha}.$$

Since $\mathbb{E}[r] = \mathbb{E}[r^*] + \rho(b^* - b)$, under independence $\|\mathbb{E}[r]\|_\infty \leq 2\alpha\rho/(n + 2\alpha)$. For the signed contrast, $d(b^*) = (1 - 2\pi_1)(1 - 2\pi_2)$ and the coefficient of each marginal shift is at most 2 in absolute value, so

$$\left|\mathbb{E}[d(r)] - 2\rho_{pq}(1 - c_{pq})\kappa_{pq}\right| \leq \frac{4\alpha\rho_{pq}}{n_{pq} + 2\alpha}.$$

These bounds describe how far the smoothed baseline moves from the exact empirical marginals; shrinkage scales the signal and this shift by the same factor.

G.3 Expressiveness of the Hyperedge Encoder

Normalization with fixed encodings. Fix the retained edges and their encodings, and write U_v for the numerator in Equation 6 and $D_v > 0$ for the sum of weights. If the scores or gates are multiplied by a common factor $\nu > 0$, positive homogeneity of ReLU gives $\Delta h_v(\nu) = \nu U_v/(\varepsilon + \nu D_v)$. For $\varepsilon = 0$ the common scale cancels; for $\varepsilon > 0$ the distance from U_v/D_v is $\varepsilon\|U_v/D_v\|/(\varepsilon + \nu D_v)$.

Joint ternary encoding. On the same input representations, a model that adds single-layer pairwise functions belongs to the class $\mathcal{P}$ of Definition E.10 and satisfies $\Delta^h_1\Delta^h_2\Delta^h_3 f = 0$, because each term misses one variable. The joint encoder can produce a nonzero mixed difference. Take scalar inputs, $W_m = W_t = W_a = 1$, and $W_r = b = 0$, and omit LayerNorm, so that $z = \mathrm{ReLU}(h_m + h_t + h_a)$. When the three inputs sum to -2 and the step is 1,

$$\Delta^{(3)}_1 z = \mathrm{ReLU}(1) - 3\,\mathrm{ReLU}(0) + 3\,\mathrm{ReLU}(-1) - \mathrm{ReLU}(-2) = 1.$$

With LayerNorm, take the two-dimensional pre-activation $(t, -t)$ with $t = h_m + h_t + h_a$, unit gain, zero bias, and $\epsilon_{\mathrm{LN}} > 0$. After ReLU, the first coordinate is $\max(t, 0)/\sqrt{t^2 + \epsilon_{\mathrm{LN}}}$, and the same difference equals $1/\sqrt{1 + \epsilon_{\mathrm{LN}}} \neq 0$. The class of joint encoders is therefore not contained in $\mathcal{P}$.

G.4 Label Use and Mean Signed Scores

Labels used in a fixed forward pass. Fix the episode index, the query batch, the parameters, the support and auxiliary labels, the observation mask, and the random state. The statistics read the meta-training columns, relation supervision on the target and the adapter use support labels of the target, and query graph construction and relation triples read the auxiliary columns. Replacing the values of the query target labels therefore leaves the prediction of this forward pass unchanged.

Mean signed scores. For auxiliary pairs that use the exact marginals of the same joint-observation distribution, the mean ideal score is $2\rho_{pq}(1-c_{pq})\kappa_{pq}$ (Appendix F.4), so scores of different property pairs are additionally scaled by $\rho_{pq}(1 - c_{pq})$.

Pairs of the target and an auxiliary property use the baseline conditioned on support labels, and their mean has to be computed separately. Let the distribution under consideration have marginals $\mu_\tau = \mathbb{E}[Y_\tau]$ and $\mu_a = \mathbb{E}[Y_a]$, and let the auxiliary marginal of the baseline be a fixed π^0_a. At every point $d(r_{\tau a}) = 2\rho_{\tau a}(2Y_\tau - 1)(Y_a - \pi^0_a)$, and taking the expectation and expanding the covariance gives

$$\mathbb{E}[d(r_{\tau a})] = 4\rho_{\tau a}\,\mathrm{Cov}(Y_\tau, Y_a) + 2\rho_{\tau a}(2\mu_\tau - 1)(\mu_a - \pi^0_a).$$

The second term vanishes when the auxiliary marginal matches π^0_a or when the target labels are balanced.

H Label Statistics Behind the Relation Model

Figure 2(a, b) averages one statistic per property pair, with equal weight on every pair that has joint observations; when the agreement expected under independence equals 1, we set $\kappa = 0$. Panels (c) to (f), and panel D of Figure 1, are the four cases in this section. Each case is one limit of supervising relations with binary agreement.

H.1 Chance Agreement under Sparse Labels: MUV-713 and MUV-733

MUV is derived from PubChem bioactivity data (Rohrer & Baumann, 2009). MUV-713 and MUV-733 are screens for inhibitors of coactivator binding to estrogen receptors α and β (PubChem AIDs 713 and 733), and 3,121 molecules are observed for both in our label matrix. Table H.1 gives the four joint states. The observed agreement is 99.8718%, against 99.8719% under independence, so $\kappa = -0.0005$. The marginals already account for it.

Table H.1: Joint label states of MUV-713 and MUV-733.

State	Observed count	Expected under independence
Both negative	3117	3117.0
Only MUV-733 positive	1	1.00
Only MUV-713 positive	3	3.00
Both positive	0	0.001

H.2 Direction of Disagreement: Increase and Decrease of the Same Readout

In the CEETOX panel of ToxCast, the assays OHPROG_up and OHPROG_dn (both with the prefix CEETOX_H295R_) record increases and decreases of the 17α-hydroxyprogesterone readout. The 500 jointly observed molecules give $\kappa = -0.3628$ (Table H.2). A binary disagreement label adds the 151 molecules with only a decrease to the 119 with only an increase. The two states are the two directions of the same readout, and the four-state target keeps them apart.

Table H.2: Joint label states of the OHPROG increase (up) and decrease (dn) assays.

	dn negative	dn positive
up negative	230	151
up positive	119	0

H.3 Molecule-Dependent Association: NR-AR and SR-p53 in Tox21

Of the 7,823 molecules in Tox21, 243 (3.1%) have a phosphorus atom in their SMILES. Table H.3 splits the jointly observed molecules of NR-AR and SR-p53 by that atom. The association is close to zero on the full set and on the molecules without phosphorus, and an order of magnitude larger on the 213 molecules that contain it. An episode-level weight for this property pair would have to use one of these three values for every molecule.

Table H.3: Cohen's κ between NR-AR and SR-p53 on subsets of Tox21.

Molecules	Joint observations	Cohen's κ
All jointly observed	6657	$+0.018$
Containing phosphorus	213	$+0.240$
Other	6444	$+0.013$

H.4 Support from Joint Observations and the Effect of a Single Label

Table H.4 relabels one joint positive $(1, 1)$ as $(1, 0)$ on two pairs with $\hat{\kappa} \approx 0.22$. The MUV pair has counts $[3058, 4, 3, 1]$, so its estimate rests on a single joint positive and falls to -0.0013 after relabeling. The PCBA pair has 1,151 joint positives among 204,894 molecules, and $\hat{\kappa}$ moves by 1.9×10^{-4}. The two sensitivities differ by a factor of about 1,200 (Figure 2f). Cohen's κ divides by $1-c_{pq}$, which is 0.0029 for the MUV pair. The C4RC target does not: the mean of its signed contrast is $2\rho_{pq}(1 - c_{pq})\kappa_{pq} = 2\rho_{pq}(a_{pq} - c_{pq})$ (Equation 3), and one relabeled molecule changes $a_{pq} - c_{pq}$ by at most $2/n_{pq}$. The sensitivity of the target therefore follows the number of joint observations, and ρ_{pq} shrinks the target further when that number is small.

Table H.4: Effect of relabeling one joint positive as $(1, 0)$.

Property pair	Joint observations	Joint positives	Original $\hat{\kappa}$	After relabeling	$\lvert\Delta\hat{\kappa}\rvert$
MUV-652, MUV-712	3066	1	$+0.2211$	-0.0013	2.2×10^{-1}
PCBA-1460, PCBA-485364	204894	1151	$+0.2231$	$+0.2230$	1.9×10^{-4}

A binary target fails each case. It counts the MUV agreement as dependence, it adds the two CEETOX states together, it gives NR-AR and SR-p53 one weight on every molecule, and it has no term that depends on how many joint observations support a relation.