DETAIL: DISENTANGLING INFORMATION TRANSFER AND OPTIMIZATION IN INTERMEDIATE-LAYER GNN DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate-layer distillation in graph neural networks (GNNs) is often said to pass useful hidden representations from a teacher to a student. But higher accuracy alone does not always show that teacher information caused the gain. The extra objective may help simply by changing how the student trains. Standard ablations cannot separate these effects because they also change the strength of the auxiliary gradient. We introduce DETAIL, a preregistered test that changes the auxiliary signal while keeping its gradient strength on the student encoder fixed. We replace the teacher's hidden states with a teacher-free variance and covariance objective or with the hidden states of an untrained network of the same architecture, and we also test random Gaussian noise in place of the auxiliary gradient. At every update, we scale each substitute's gradient on the student encoder to the same norm as the real teacher's. On `ogbn-arxiv`, the intermediate objective improves a GraphSAGE student with 16 hidden units per layer by +1.04 percentage points (pp; 95% confidence interval [0.77, 1.32]) over a baseline that already uses logit distillation. The teacher-free objective and the untrained network recover 0.98 and 0.96 of this gain, with a $+0.13$ pp upper confidence bound on the real teacher's added benefit over the teacher-free objective. Random noise of the same size recovers 0.76 of the gain on its own. A standard per-dimension objective and a student with 64 hidden units give the same picture. On `ogbn-Mag`, a rule fixed before any run with the objective predicted that it would help. It did (+0.83 pp [0.59, 1.07]), and the substitutes again recovered most of the gain (0.88 and 0.95). On `ogbn-Products`, the largest graph, the gain is $+0.13$ pp $[-0.03, 0.29]$, even with the same exposure as on `ogbn-arXiv`. Within the tested settings, most of the gain can be reproduced without learned intermediate teacher representations.

1 INTRODUCTION

Knowledge distillation trains a student model using supervision from a stronger teacher (Hinton et al., 2014; Zagoruyko & Komodakis, 2017; Buciluǎ et al., 2006). In standard distillation, the student learns from the teacher's output predictions. Intermediate-layer distillation additionally encourages the student to reproduce aspects of the teacher's internal representations (Romero et al., 2015). When these objectives improve accuracy, the improvement is commonly attributed to useful information encoded in these hidden representations (Tian et al., 2025). But higher accuracy alone does not show that the student learned useful information from the teacher's hidden representations. The intermediate distillation loss also changes how the student trains, which may itself improve accuracy. The gain could therefore come from teacher information, this training effect, or both. Common ablations remove or replace the teacher target, but doing so also changes the gradient produced by the auxiliary loss. They therefore cannot cleanly separate these effects.

This issue is especially important in Graph Neural Networks (GNNs) (Yang et al., 2020; Tian et al., 2025; Joshi et al., 2024), where intermediate objectives depend on several factors like node correspondence, graph structure, layer routing, representation geometry, and loss scale. Replacing a teacher representation with a control target can change not only the information presented to the student but also the magnitude of the resulting auxiliary gradient; differences in gradient scale alone can substantially alter multi-objective training dynamics (Chen et al., 2018). More broadly, prior work

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.

has shown that distillation can improve students even with weak or randomly initialized teachers, suggesting that gains need not arise solely from information encoded by a trained teacher (Yuan et al., 2020; Sarnthein et al., 2023). A clean test must therefore vary the source of the intermediate target while controlling the strength of the resulting optimization intervention.

Following FitNets (Romero et al., 2015), we call such an intermediate target a *hint*, and the extra loss that uses it an *auxiliary objective*. Recent GNN distillation methods often supervise students using the teacher's internal representations in addition to its output predictions. These objectives may align node embeddings, local topology, global structural relationships, or combinations of such signals (Yang et al., 2020; Joshi et al., 2024; He et al., 2022; Yang et al., 2022). We study this question using a multi-layer structural and embedding hint (Ahluwalia et al., 2025): *do its gains come from information learned by the teacher, or from the optimization effect of the auxiliary objective?* The same ambiguity appears whenever student training includes an intermediate teacher target, so our test can also be applied to other representation-distillation methods.

We do not propose a new distillation method. Instead, we introduce DETAIL, a test that separates teacher information from the optimization effect of the intermediate objective using gradient-matched source substitution. Starting from the real-teacher hint, we compare it with three controls that do not use learned intermediate teacher representations. The *first* uses hidden states from an architecture-matched network that is never trained. The *second* removes the teacher target entirely and applies a variance-covariance objective to the student's own representations. The *third* applies random Gaussian noise directly to the student encoder gradient. For every seed and training update, we scale each control so that its encoder-gradient norm matches that of the real-teacher hint. This controls the magnitude of the optimization intervention while changing the source and structure of the intermediate supervision.

We add two further controls. First, the node-permutation control shuffles the correspondence between teacher and student nodes. It tests whether the correct node pairing matters. Second, the auxiliary-removal experiment applies the auxiliary objective only during the first 150 of 3,000 training updates, then removes it and continues ordinary training.

Measuring what remains after removal shows how much of the early advantage survives ordinary training, and whether the size of that early advantage predicts the final accuracy.

Our main contributions are summarized as follows:

- We identify a central ambiguity in intermediate GNN distillation. Standard ablations typically change both the information provided by the teacher and the optimization effect of the auxiliary loss. As a result, an accuracy gain alone cannot show which factor caused the improvement.
- We introduce DETAIL, a preregistered source-substitution framework that changes the source of the intermediate target while matching the real-teacher encoder-gradient norm for every seed and update. Correspondence, matched-noise, and auxiliary-removal controls help separate the effects of learned teacher information, target structure, and optimization pressure.
- We show that most of the observed benefit can be reproduced without learned intermediate teacher representations. On `arXiv`, the main hint improves a width-16 student by $+1.04$ percentage points, yet teacher-free and untrained-network controls recover nearly all of this gain. Matched noise recovers most of the gain, while correct node pairing still matters for the representation-based objective. Similar substitution results hold on the paper-citation graph of `Mag`. On `Products`, we find no clear gain even after matching the auxiliary-to-task gradient ratio.

2 Background

2.1 Setting and output-level baseline

Let $G = (V, E, X)$ denote a graph. A fixed teacher f_T and a trainable student f_S produce logits z^T and z^S, with hidden representations $\mathcal{H}^T = \{H_1^T, \ldots, H_{L_T}^T\}$ and $\mathcal{H}^S = \{H_1^S, \ldots, H_{L_S}^S\}$. The main baseline combines cross-entropy with output-level distillation:

$$\mathcal{L}_{\text{task}} = (1-\alpha)\mathcal{L}_{\text{CE}} + \frac{\alpha\tau^2}{C}\,\mathrm{KL}_{\text{batchmean}}\left(\text{softmax}(z^T/\tau) \,\middle\|\, \text{softmax}(z^S/\tau)\right), \tag{1}$$

where C is the number of classes and τ is the distillation temperature. We use $\alpha = 0.5$ and $\tau = 4$. We denote the resulting output-distillation baseline by **B**. To match the computation used in the intermediate-distillation conditions, **B** also executes the auxiliary forward pass, but assigns zero weight to the intermediate loss. Thus, all effects of the intermediate objective are measured relative to a student that already receives output-level distillation.

2.2 Intermediate-Layer Distillation Objective

We test a published multi-layer hint (Ahluwalia et al., 2025). It uses layer-specific projections, an embedding dissimilarity and the G-CRD structural loss (Joshi et al., 2024). For each teacher layer ℓ and student layer m, the hint combines an embedding loss and a structural loss: $M_{\ell m} = (D_{\mathrm{emb}})_{\ell m} + (D_{\mathrm{str}})_{\ell m}$. The embedding term measures disagreement between the projected teacher and student representations, while the structural term contrasts each student node with its matching teacher node and the other sampled teacher nodes (Joshi et al., 2024). Appendix A.1 gives the exact implementation.

The original method learns row-normalized attention over layer pairs. In our main experiments, however, all teacher–student layer pairs are weighted equally. Specifically, the implementation uses

$$A^{\mathrm{unif}}_{\ell m} = \frac{1}{L_T L_S}, \qquad \mathcal{L}_{\mathrm{aux}} = \mathbf{1}_{L_T}^{\top}(A^{\mathrm{unif}} \odot M)\frac{\mathbf{1}_{L_S}}{L_S}. \tag{2}$$

This is equivalent to $\mathrm{mean}(M)/L_S$. The extra factor $1/L_S$ comes from the final averaging over student layers. Therefore, this implementation is L_T times smaller than a row-stochastic uniform weighting with entries $1/L_S$; when $L_T = 3$, its magnitude is one third as large. Because A^{unif} is fixed, the learned attention path is not used. Our main experiments therefore test the uniformly weighted hint rather than learned routing.

3 Methodology

An accuracy gain from the intermediate hint does not, by itself, prove that the gain comes from the teacher's learned representations. The auxiliary gradient also changes the student's optimization trajectory. DETAIL tests whether the same gain can be recovered without learned intermediate teacher representations. We compare the trained-teacher condition (CC) with three controls: hidden states from an untrained teacher of the same architecture (CT), a variance–covariance loss on the student's own features (GV), and isotropic Gaussian noise added directly to the encoder gradient (RN) (Table 1). For each seed r and update $u \in \{1, \ldots, 150\}$, we rescale each control's auxiliary encoder gradient so that

$$\left\|\widetilde{g}^{s}_{r,u}\right\|_2 = \left\|g^{\mathrm{CC}}_{r,u}\right\|_2, \qquad s \in \{\mathrm{CT}, \mathrm{GV}, \mathrm{RN}\}, \tag{3}$$

where $g^{\mathrm{CC}}_{r,u}$ is recorded in the trained-teacher run with the same seed and student width. RN tests how much of the gain can be recovered by a perturbation of the same magnitude alone. CT and GV test whether structured interventions that do not use learned teacher representations recover more of the gain. Their differences from CC bound the additional benefit associated with the trained teacher. We match gradient norms before clipping; gradient directions, their alignment with the task gradient, and the resulting optimizer updates can still differ.

3.1 Source substitution

We compare the baseline B with the trained teacher condition CC and three controls: CT, GV, RN (Table 1). The trained teacher provides output-level distillation in all five conditions. The substitutions change only the intermediate supervision or its gradient; they do not remove the teacher's output predictions.

Variance and covariance control. GV uses the student's own features without a teacher target. Let $\mathcal{N} \subseteq V_{\mathrm{train}}$ contain the n sampled training nodes and let

$$Z = H^{S}_{L_S}[\mathcal{N}] \in \mathbb{R}^{n \times d} \tag{4}$$

Table 1: Labels used throughout the paper.

Label	Meaning
Baselines	
● B	Cross-entropy plus logit distillation; no intermediate hint
B_0	Cross-entropy only; no logit or intermediate distillation
Real-teacher hint	
■ CC	Real-teacher hint with correct node correspondence (reference)
× CP	Real-teacher hint with projector targets permuted across nodes
PC, PP	Encoder targets permuted (PC), or both encoder and projector targets permuted (PP)
Teacher-independent controls matched to CC in encoder-gradient norm	
♦ CT	Hint from an untrained network with the teacher architecture
● GV	Teacher-free variance–covariance objective on student features
▲ RN	Isotropic Gaussian noise applied directly to the encoder gradient
Per-dimension hint	
□ H2CC, ◇ H2CT, ○ H2GV	CC, CT, and GV under the per-dimension FitNets-style objective
Dose-matched Products	
CC_ρ, GV_ρ, B'	`Products` runs using the `arXiv` exposure schedule; B' is the corresponding baseline

denote their raw representations at the final student encoder layer. We apply the variance and covariance terms of VICReg (Bardes et al., 2022):

$$\mathcal{L}_{\mathrm{var}}(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left\{0,\, 1 - \sqrt{\mathrm{Var}(Z_{:j}) + \epsilon}\right\}, \tag{5}$$

$$\mathcal{L}_{\mathrm{cov}}(Z) = \frac{1}{d}\sum_{\substack{1\le p,q\le d\\ p\neq q}} \mathrm{Cov}(Z)_{pq}^2, \tag{6}$$

$$\mathcal{L}_{\mathrm{GV}}(Z) = \gamma_{\mathrm{var}}\mathcal{L}_{\mathrm{var}}(Z) + \gamma_{\mathrm{cov}}\mathcal{L}_{\mathrm{cov}}(Z). \tag{7}$$

Variance and covariance are computed across the sampled nodes. The variance term penalizes feature standard deviations below 1, while the covariance term penalizes off-diagonal covariances. GV is applied once to the final representation, rather than separately to each teacher-student layer pair. The weights γ_{var} and γ_{cov}, and the numerical stabilizer $\epsilon > 0$, are fixed. The gradient-matching rule determines the overall scale at each update. Appendix A.1 reports how the two terms behave on each dataset.

Per-dimension regression. The embedding term in Section 2.2 compares feature averages, so it does not directly align individual coordinates. To check whether our findings depend on this reduction, we also test a FitNets-style regression objective, denoted H2 (Romero et al., 2015). For each teacher layer ℓ and student layer m, a learned linear regressor $P_{\ell m}$ maps the student's representation to the teacher's $d_T = 384$ features. The loss is

$$\mathcal{L}_{\mathrm{H2}} = \frac{1}{L_T L_S}\sum_{\ell=1}^{L_T}\sum_{m=1}^{L_S} \frac{\left\|P_{\ell m}\big(H_m^S[\mathcal{N}]\big) - H_\ell^T[\mathcal{N}]\right\|_F^2}{n d_T}, \tag{8}$$

where $P_{\ell m}$ is applied to each node representation. The loss averages squared errors over nodes, feature coordinates, and all $L_T L_S = 6$ layer pairs, using raw, unnormalized features.

H2CC uses the trained teacher's hidden states. H2CT replaces them with hidden states from the untrained teacher, while H2GV replaces the regression loss with $\mathcal{L}_{\mathrm{GV}}$. Sampling, routing, the training schedule, the optimizer reset, and paired seeds are unchanged. H2CC follows the original CC run's ratio of auxiliary to task gradient norms. H2CT and H2GV are then matched to H2CC's encoder gradient norm for the same seed and update. Appendix A.6.1 gives further details.

3.2 Gradient-matched exposure

Let $\theta^s_{E,r,u}$ denote the student encoder parameters in condition s, for seed r, immediately before update u. For conditions defined by an auxiliary loss, the raw auxiliary gradient is

$$g^s_{r,u} = \nabla_{\theta_E}\mathcal{L}^s_{\mathrm{aux}}\big|_{\theta_E=\theta^s_{E,r,u}}. \tag{9}$$

Each gradient is evaluated along its condition's own training trajectory. For each seed and student width, we first record the CC gradient norms over updates 1–150. This reference trace is held fixed during the replacement runs. For a nonzero replacement gradient, we set

$$\lambda^s_{r,u} = \frac{\|g^{\mathrm{CC}}_{r,u}\|_2}{\|g^s_{r,u}\|_2}, \qquad \widetilde{g}^s_{r,u} = \lambda^s_{r,u} g^s_{r,u}. \tag{10}$$

The coefficient $\lambda^s_{r,u}$ is used only to rescale the computed gradient.

For RN, there is no auxiliary loss. Instead, we draw $g^{\mathrm{RN}}_{r,u} \sim \mathcal{N}(0, I_{p_E})$, where p_E is the number of encoder parameters, and apply Eq. (10). Thus, CT, GV, and RN receive the same auxiliary gradient norm as CC at each update. A zero reference norm requires a zero applied gradient. If the replacement gradient is zero while the reference norm is positive, exact matching cannot be achieved by rescaling.

Transferring the exposure schedule. For the dose-matched `Products` experiment and H2CC, we use the auxiliary-to-task gradient ratio from the frozen `arXiv` width-16 reference trace:

$$\rho(u) = \frac{\|g^{\mathrm{ref}}_{\mathrm{aux},u}\|_2}{\|g^{\mathrm{ref}}_{\mathrm{task},u}\|_2}. \tag{11}$$

Both gradients are measured on the student encoder before clipping. Let $g^s_{\mathrm{task},r,u}$ denote the task gradient at the current parameters of condition s. We rescale its auxiliary gradient to satisfy

$$\|\widetilde{g}^s_{r,u}\|_2 = \rho(u)\, \|g^s_{\mathrm{task},r,u}\|_2. \tag{12}$$

The auxiliary gradient therefore follows the reference ratio while adapting to the task-gradient magnitude in the current run. We apply this rule to both CC and GV in the dose-matched `Products` experiment, and to H2CC. H2CT and H2GV then follow Eq. (10), using the H2CC trace as their reference.

Both matching rules control the auxiliary encoder-gradient norm before clipping. They leave its direction and its alignment with the task gradient unconstrained. Consequently, the combined gradient after clipping and the resulting Adam update can still differ across conditions. Gradients of parameters used only by the auxiliary objective are outside the matching constraint. Appendix A.1.2 reports the matching errors and scale diagnostics.

3.3 Correspondence controls

Source substitution asks whether the target must come from a trained teacher. We separately test whether that target must be assigned to the correct node. The hint contains two target paths that can be permuted independently. The *encoder target* determines which teacher node representation is paired with each student node in the part of the auxiliary objective that backpropagates through the student encoder. The *projector target* determines the teacher-node assignment used by the trainable auxiliary projector. We cross these two assignments in a 2×2 factorial. For each path, C denotes correct node correspondence and P denotes a fixed permutation across nodes, giving CC, CP, PC, and PP. For example, CP keeps the encoder target correctly aligned with each student node but permutes the teacher-node assignment used by the projector. The paired interaction is

$$\Delta_{\mathrm{int}} = (\mathrm{CC} - \mathrm{CP}) - (\mathrm{PC} - \mathrm{PP}). \tag{13}$$

We also run a second correspondence test with a single target assignment. It compares a correctly aligned target S, a permuted target S_{perm}, and a permuted target whose encoder-gradient norm is matched to S, denoted $S_{\mathrm{perm_match}}$. These controls test whether the gain depends on correct node assignment rather than only on the training schedule, the extra forward pass, or the presence of an additional gradient.

3.4 AUXILIARY REMOVAL AND MEASURED EFFECTS

The auxiliary objective is active for the first 150 updates. We then remove it, reset the Adam optimizer in every condition, including **B**, and continue training with $\mathcal{L}_{\text{task}}$ for another 2,850 updates. Applying the same reset to every condition prevents the reset itself from explaining differences between conditions. Our primary endpoint is test accuracy at update 3,000.

For condition X and paired seed r, we define the endpoint effect relative to the baseline as

$$\delta_r^X = \text{Acc}_r^X(3000) - \text{Acc}_r^{\mathbf{B}}(3000). \tag{14}$$

To measure how much of the real-teacher hint's gain is recovered by a substitute X, we define

$$R^X = \frac{\bar{\delta}^X}{\bar{\delta}^{\text{CC}}}, \qquad \bar{\delta}^X = \frac{1}{N}\sum_{r=1}^{N} \delta_r^X, \tag{15}$$

where $\bar{\delta}^X$ denotes the mean paired effect over seeds. We compute R^X only when the preregistered CC–B gate is satisfied, which requires the 95% confidence interval of CC $-$ B to lie above zero. This avoids interpreting a ratio whose denominator is too small to represent a reliable real-teacher gain.

We also compare CC directly with each substitute. For example, the upper confidence limit of $\text{CC} - \text{GV}$ bounds the additional accuracy benefit supported by the data for using the trained teacher's intermediate representations rather than the GV objective. It does not measure the teacher's total contribution, since output-level distillation from the same teacher remains active in both conditions. No equivalence margin was preregistered. Therefore, failure to distinguish a substitute from CC does not establish that the two objectives are equivalent. We report each recovery fraction with a 95% seed-keyed paired-bootstrap interval: 20,000 resamples for GV and CT, and 10,000 for RN and H2, following their respective preregistrations.

For a structured substitute $X \in \{\text{CT}, \text{GV}\}$, the real-teacher gain can also be decomposed as

$$\text{CC} - B = (\text{RN} - B) + (X - \text{RN}) + (\text{CC} - X) \tag{16}$$

The three terms measure, respectively, the effect recovered by a same-magnitude random perturbation, the additional effect of a structured teacher-independent intervention, and the additional effect of the trained-teacher intermediate hint relative to that substitute. Finally, we track accuracy after the auxiliary objective is removed. These trajectories show whether an advantage created during the first 150 updates persists during subsequent training with $\mathcal{L}_{\text{task}}$ alone. The endpoint comparisons follow the preregistered analysis plan; the trajectory analysis is descriptive and post hoc, although the checkpoint grid was fixed before this analysis.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Datasets and models. We use the official OGB splits for ogbn-arxiv, ogbn-products and ogbn-mag (Hu et al., 2020), which we call `arXiv`, `Products` and `Mag`. For `Mag`, we use only the paper nodes and citation edges (Appendix A.2). All three graphs are symmetrized and have no added self-loops. Students are two-layer GraphSAGE models (Hamilton et al., 2017) with 16 or 64 hidden units per layer, which we call width 16 and width 64. Teachers are three-layer GATv2 models with 128 hidden units and three heads (Brody et al., 2022). Table 2 gives the train, validation, and test accuracy of each teacher and of the baseline student B. Unless stated otherwise, accuracy effects are measured relative to baseline B. Appendix A.2 gives the full dataset and checkpoint details.[1]

Training protocol. All main runs are full batch and follow the schedule of Section 3.4. Appendix A.1.1 lists every hyperparameter and how each was chosen.

Evaluation method. Each main cohort has 30 paired seeds. The endpoint is test accuracy at update 3,000, not best-validation accuracy. We report paired mean differences in percentage points (pp) with two-sided Student-t 95% confidence intervals. Historical experiments use $n = 3$ or $n = 5$ and are not pooled with the main cohorts. Table 6 gives the provenance of every result.

[1]We will release the code and experimental records upon acceptance.

Table 2: Absolute accuracy (%) of each teacher and of the baseline student B (fixed-final mean over 30 seeds), with the teacher–student gaps. Every result in Section 4 is measured relative to baseline B. `Products` uses a teacher cache built over training rows only.

Model	Train	Valid	Test	Valid − test
arXiv teacher	73.64	72.14	71.19	0.95
arXiv student B, width 16	66.15	66.00	65.18	0.83
arXiv student B, width 64	72.73	70.34	69.35	0.98
arXiv gap, width 16	+7.48	+6.14	+6.01	—
Products teacher	91.40	90.42	80.21	10.21
Products student B, width 16	90.56	89.21	71.27	17.94
Products gap	+0.85	+1.21	+8.94	—
Mag teacher	31.61	33.37	33.73	−0.36
Mag student B, width 16	28.16	28.45	29.76	−1.31
Mag gap	+3.46	+4.91	+3.97	—

4.2 Routing audit

We first reproduced the comparison that motivated this study (Appendix A.3). In our runs at the published setting ($\alpha = 0.9$, $T = 4$), the logit-distillation baseline scored below plain cross-entropy. With a corrected baseline ($\alpha = 0.5$), uniform routing matched or beat learned routing in all six run directories. Our reproduced student is weaker than the published one and each result uses 3 runs, so this checks one configuration and makes no claim about learned routing in general. A cross-entropy-only baseline, B_0 (B with $\alpha = 0$), run with 30 paired seeds, finds no detectable difference between the corrected baseline and plain cross-entropy (Section 4.3).

4.3 Main result on arXiv

The real-teacher hint improves both students over B: $CC - B$ is $+1.04$ pp $[+0.77, +1.32]$ at width 16, positive in 28 of 30 seeds, and $+0.36$ pp $[+0.30, +0.42]$ at width 64, positive in 29 of 30 seeds (Table 7). The effect is about three times larger at width 16. We detect no accuracy difference between logit distillation and plain cross-entropy here. Over 30 paired seeds, $B - B_0$ is -0.04 pp $[-0.13, +0.05]$ at width 16 and $+0.01$ pp $[-0.04, +0.05]$ at width 64.

4.4 Controls and gradient-matched source substitution

4.4.1 Node correspondence and projector targets

Destroying node correspondence is strongly harmful. Relative to the correctly aligned arm, a permuted target gives -2.3975 pp $[-2.7039, -2.0912]$, and a gradient-matched permuted target gives -1.9853 pp $[-2.3257, -1.6449]$; both are negative in 30 of 30 seeds. In the factorial experiment, permuting the encoder target changes CC by -2.0152 pp $[-2.3791, -1.6513]$ at width 16 and -0.7024 pp $[-0.7898, -0.6149]$ at width 64. By contrast, with the encoder target correct, permuting the projector target changes accuracy by only -0.0387 pp $[-0.1563, +0.0789]$ and $+0.0086$ pp $[-0.0557, +0.0730]$. Thus, a wrongly assigned encoder target is harmful, while the projector assignment has no measurable effect. These controls also show that the schedule, the reset, and the extra forward pass do not explain the gain; Section 4.4.2 shows how much of it random noise explains. Table 8 reports every registered contrast, and Table 9 reports the full interaction.

4.4.2 Gradient-matched source substitution

Table 3 is the core confirmatory result. GV removes the teacher hidden states and uses a variance–covariance objective. CT uses hidden states from a never-trained architecture-matched network. Each substitute matches the real-teacher CC encoder-gradient norm for every seed and update. Real-teacher logit distillation remains active in every arm.

The teacher-free objective recovers 0.983 $[0.877, 1.090]$ of the width-16 gain, and the untrained twin recovers 0.956 $[0.863, 1.056]$. At width 64, the fractions are 1.001 $[0.832, 1.191]$ and 0.915 $[0.779, 1.065]$. The upper confidence bound on what the real teacher adds beyond GV is $+0.13$ pp at width 16 and $+0.07$ pp at width 64. These are confidence bounds, not equivalence claims: no equivalence margin was preregistered.

The noise arm RN recovers 0.765 [0.634, 0.897] of the gain and is 0.25 pp [0.09, 0.40] below CC. So noise of the same size gives most, but not all, of the gain. The per-dimension hint H2 gains +1.07 pp,

Table 3: Source substitution. $X - B$ is the gain over the baseline; $X - \mathrm{CC}$ bounds the real teacher's added contribution. At width 64, all arms come from one cohort whose CC − B is $+0.353$ pp; Appendix Table 7 reports $+0.36$ pp from a separate cohort.

Cohort	Arm	$X - B$	$X - \mathrm{CC}$ (bound)	Fraction of CC − B
arXiv, width 16	GV	$+1.0244$ $[+0.7361, +1.3126]$	-0.0182 $[-0.1316, +0.0953]$	0.983 [0.877, 1.090]
arXiv, width 16	CT	$+0.9969$ $[+0.7322, +1.2616]$	-0.0456 $[-0.1506, +0.0594]$	0.956 [0.863, 1.056]
arXiv, width 64	GV	$+0.3533$ $[+0.2841, +0.4224]$	$+0.0003$ $[-0.0660, +0.0667]$	1.001 [0.832, 1.191]
arXiv, width 64	CT	$+0.3230$ $[+0.2635, +0.3825]$	-0.0299 $[-0.0848, +0.0250]$	0.915 [0.779, 1.065]
arXiv, width 16	GV − CT	—	$+0.0274$ $[-0.0969, +0.1518]$	—
arXiv, width 16	RN	$+0.7973$ $[+0.5413, +1.0533]$	-0.2453 $[-0.4002, -0.0903]$	0.765 [0.634, 0.897]
Mag, width 16	GV	$+0.7260$ $[+0.5114, +0.9405]$	-0.1021 $[-0.2128, +0.0087]$	0.877 [0.763, 1.005]
Mag, width 16	CT	$+0.7873$ $[+0.5796, +0.9949]$	-0.0408 $[-0.1551, +0.0736]$	0.951 [0.832, 1.094]

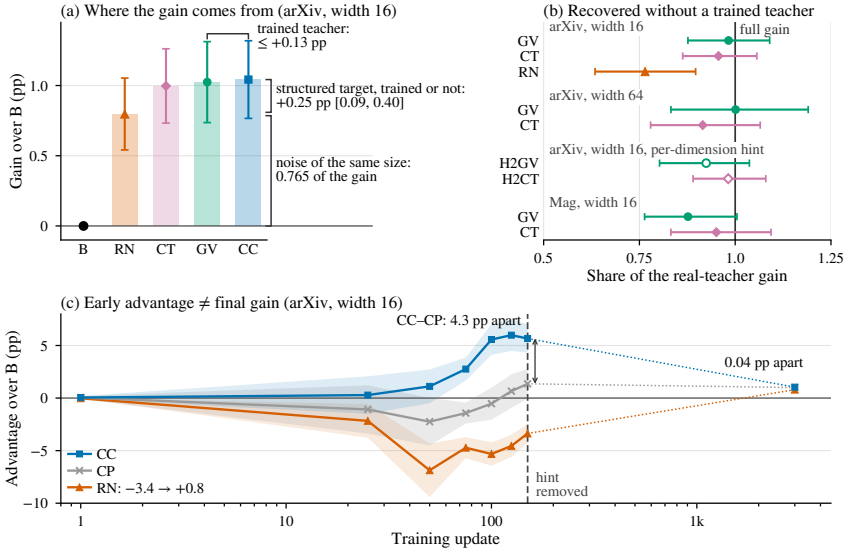


Figure 1: Gradient-matched source substitution (top) and paired test-accuracy differences from each cohort baseline (bottom). Error bars and bands show 95% intervals over 30 paired seeds. The vertical line marks removal of the auxiliary objective at update 150; Appendix Tables 10 and 11 give the values behind Figure (c).

and its untrained and teacher-free substitutes recover 0.98 and 0.92 of it, so the small teacher share is not specific to one hint (Figure 1b; Appendix A.6.1).

4.4.3 Switch-off analysis

The auxiliary stops at update 150, but the confirmatory endpoint is update 3,000. At width 16, the real-teacher advantage is $+5.66$ pp $[4.31, 7.02]$ at switch-off and $+1.043$ pp at the endpoint, so only 18.4% remains after 2,850 ordinary updates. At width 64, $+1.87$ pp falls to $+0.353$ pp, or 18.8%. The same broad decay appears for every correctly aligned source. The source-specific differences are much larger at switch-off than at the endpoint. For example, CC and CP are 4.31 pp apart at width 16 when the auxiliary ends but only 0.039 pp apart at update 3,000. The noise arm shows the same without any target: it is 3.4 pp behind the baseline when the auxiliary is removed, yet ends 0.8 pp ahead (Appendix Table 11). So the early advantage is not an early version of the final gain. It does not imply that every arm retains a fixed fraction: the full trajectories include sign changes and different peak times. Figure 1c shows the full decay. The trajectory analysis is descriptive and post hoc; Table 10 reports every checkpoint.

4.5 Cross-dataset test on Products and Mag

Table 4 reports the registered natural-exposure test and a dose-matched follow-up. At natural exposure, the CC–B gate, which here also requires an estimate of at least 0.30 pp, does not pass, so no recovery fraction is computed. The natural Products exposure ratio is 0.220, compared with 0.736 on arXiv. We therefore repeat the test while imposing the frozen arXiv width-16 exposure trajectory at every update. The delivered integrated encoder norm is 7.666, versus 7.468 on arXiv, a difference of $+2.7\%$.

Table 4: Fixed-final Products results in pp with 95% confidence intervals. The second block matches the arXiv exposure trajectory. A $\leq$ entry is the upper limit of the 95% confidence interval.

Setting and contrast	Estimate	Verdict
Natural: CC − B	$+0.0147$ $[-0.1292, +0.1586]$, 14/30	Gate not passed; $\leq +0.16$ pp
Natural: GV − B	-0.0024 $[-0.0788, +0.0740]$, 16/30	$\leq +0.07$ pp
Natural: GV − CC	-0.0171 $[-0.1526, +0.1183]$	Bound only
Dose matched: $CC_\rho - B'$	$+0.1288$ $[-0.0326, +0.2901]$, 20/30	$\leq +0.29$ pp
Dose matched: $GV_\rho - B'$	-0.0362 $[-0.1640, +0.0916]$, 15/30	$\leq +0.09$ pp
Dose matched: $GV_\rho - CC_\rho$	-0.1650 $[-0.3315, +0.0015]$, $p = 0.052$	Bound only

Table 5: Where the hint helps. Baseline progress is the baseline's training accuracy at update 150 as a share of its final training accuracy (Table 12). Teacher lead is teacher minus student validation accuracy (Table 2). The Mag row was predicted before any hinted run. Teacher adds $\leq$ is the upper 95% confidence limit of CC − GV.

Cohort	Progress at 150	Teacher lead (pp)	CC − B (pp)	GV, CT recovery	Teacher adds $\leq$
arXiv, width 16	72.0%	6.14	$+1.04$ $[+0.77, +1.32]$	0.98, 0.96	$+0.13$
arXiv, width 64	82.8%	—	$+0.36$ $[+0.30, +0.42]$	1.00, 0.92	$+0.07$
Products	95.4%	1.21	$+0.015$ $[-0.13, +0.16]$	not computed	—
Products, arXiv exp.	95.4%	1.21	$+0.13$ $[-0.03, +0.29]$	not computed	—
Mag	61.8%	4.91	$+0.83$ $[+0.59, +1.07]$	0.88, 0.95	$+0.21$

At natural strength, the real-teacher effect is $+0.015$ pp $[-0.13, +0.16]$. Under arXiv-matched exposure, it is $+0.13$ pp $[-0.03, +0.29]$, not the $+1.04$ pp arXiv effect. Thus, under-dosing does not fully explain the cross-dataset difference, although the interval allows a small effect. Two observations explain why little acceleration is available. By update 150, the Products baseline has reached 95.4% of its final training accuracy, compared with 72.0% for arXiv width 16. Its switch-off head start is only $+0.19$ pp, rather than $+5.66$ pp. Also, the Products teacher leads the student by only 0.85 pp on the training rows used by the hint, although the test gap is 8.94 pp. The auxiliary does not directly observe most of that test deficit. These observations are descriptive, but they are consistent with the small final bound.

Mag: a gain predicted in advance. Before any hinted run on Mag, a preregistered rule said the hint should help if the baseline had reached at most 80% of its final training accuracy by update 150 and the teacher led the student by at least 3.0 pp on validation. Three pilot runs gave 61.8% and 5.73 pp, so the rule predicted a gain. In the full cohort the values are 61.8% and 4.91 pp (Table 5). It held: CC $- B$ is $+0.83$ pp $[+0.59, +1.07]$, positive in 29 of 30 seeds. The substitutes again reproduce most of the gain (Table 3), and the upper confidence limit on what the trained teacher adds is $+0.21$ pp. As on arXiv, the gaps between arms at removal mostly close by the end (Appendix Table 11).

Appendix A reports the full trajectory table, endpoint split analysis, historical window and capacity checks, teacher probes, negative results on snap-patents and roman-empire, and all reproducibility and compute checks.

5 Conclusion

DETAIL tests whether the gain from an intermediate objective needs learned teacher information. On ogbn-arxiv, a uniform hint improves a logit-distilled student by $+1.04$ pp $[+0.77, +1.32]$ at width 16 and $+0.36$ pp $[+0.30, +0.42]$ at width 64. Yet a teacher-free objective and a never-trained network recover nearly all of this gain. The data bound the learned teacher's added contribution at $+0.13$ pp and $+0.07$ pp, respectively. Two things about the target still matter: permuting which node gets which target costs about 2 pp, and noise of the same size recovers only 0.76 of the gain. For CC, a $+5.66$ pp advantage at switch-off falls to 18.4% after 2,850 ordinary updates, and the noise arm ends ahead despite an early deficit, so an early advantage is not needed for the final gain. On ogbn-products, the effect is $+0.015$ pp $[-0.13, +0.16]$ at natural exposure and $+0.13$ pp $[-0.03, +0.29]$ under matched exposure, when the baseline has already reached 95.4% of its final training accuracy. Within the tested objectives, students, and datasets, most of the gain does not need learned teacher representations, although the real-teacher hint still outperforms noise of the same size. These are confidence bounds, not equivalence claims, and they do not cover learned routing or all GNN distillation methods.

AI USE STATEMENT

We used AI tools to assist with manuscript preparation, including grammar checking, readability improvements, and the generation of selected figures based exclusively on data extracted from our own experiments. We did not use generative AI tools to generate data, develop theoretical claims or proofs, design the methodology or experiments, implement the method, perform experiments, or interpret the experimental results. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work does not involve human subjects, private or personally identifiable data, or deployment in a high-stakes decision-making setting. We are not aware of direct ethical concerns beyond the standard risks associated with the misuse of machine learning methods. The authors have read and agree to adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing the model, training, and evaluation details in Sections 3, 4, and Appendix A. We will provide source code and scripts for reproducing all the results upon acceptance of the paper.

REFERENCES

Anshul Ahluwalia, Payman Behnam, Rohit Das, Alind Khare, Biswadeep Chakraborty, Pan Li, and Alexey Tumanov. STRIDE: Structure and embedding distillation with attention for graph neural networks, 2025. arXiv:2310.15938v2.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. *International Conference on Learning Representations (ICLR)*, 2022.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *International Conference on Learning Representations (ICLR)*, 2022.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *International Conference on Machine Learning (ICML)*, 2018.

Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S. Dhillon. Node feature extraction by self-supervised multi-scale neighborhood prediction. *International Conference on Learning Representations (ICLR)*, 2022.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. *International Conference on Computer Vision (ICCV)*, 2019.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Neural Information Processing Systems (NeurIPS)*, 2017.

Huarui He, Jie Wang, Zhanqiu Zhang, and Feng Wu. Compressing deep graph neural networks via adversarial knowledge distillation. *Proceedings of the Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*, 2014.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Neural Information Processing Systems (NeurIPS)*, 2020.

Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. *International Conference on Learning Representations (ICLR)*, 2021.

Chaitanya K. Joshi, Fayao Liu, Xu Xun, Jie Lin, and Chuan-Sheng Foo. On representation knowledge distillation for graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 35:4656–4667, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser-Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Neural Information Processing Systems (NeurIPS)*, 2021.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? *International Conference on Learning Representations (ICLR)*, 2023.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. *International Conference on Learning Representations (ICLR)*, 2015.

Felix Sarnthein, Gregor Bachmann, Sotiris Anagnostidis, and Thomas Hofmann. Random teachers are good teachers. *International Conference on Machine Learning (ICML)*, 2023.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V. Chawla. Learning MLPs on graphs: A unified view of effectiveness, robustness, and efficiency. *International Conference on Learning Representations (ICLR)*, 2023.

Yijun Tian, Shichao Pei, Xiangliang Zhang, Chuxu Zhang, and Nitesh V. Chawla. Knowledge distillation on graphs: A survey. *ACM Computing Surveys*, 57(8):1–16, 2025.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *International Conference on Learning Representations (ICLR)*, 2020.

Lirong Wu, Yunfan Liu, Haitao Lin, Yufei Huang, and Stan Z. Li. Teach harder, learn poorer: Rethinking hard sample distillation for GNN-to-MLP knowledge distillation. *International Conference on Information and Knowledge Management (CIKM)*, 2024.

Chenxiao Yang, Qitian Wu, and Junchi Yan. Geometric knowledge distillation: Topology compression for graph neural networks. *Neural Information Processing Systems (NeurIPS)*, 2022.

Ling Yang, Ye Tian, Minkai Xu, Zhongyi Liu, Shenda Hong, Wei Qu, Wentao Zhang, Bin Cui, Muhan Zhang, and Jure Leskovec. VQGraph: Rethinking graph representation space for bridging GNNs and MLPs. *International Conference on Learning Representations (ICLR)*, 2024.

Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old MLPs new tricks via distillation. *International Conference on Learning Representations (ICLR)*, 2022.

A Additional Experimental Details and Results

The main paper reports the central tests. This appendix preserves the complete experimental record, including full tables, historical analyses, diagnostics, and integrity checks.

A.1 Implementation and training details

The uniform-routing branch returns a constant matrix and disconnects the learned attention path from the computation graph, so the thirty-one attention-path tensors receive no gradient; two node-importance terms remain connected to the structural loss. The substitution test is therefore defined only for constant, gradient-free routing and does not test learned routing.

For the teacher-free GV objective, the minimum baseline feature standard deviation on `Products` is 2.27 at update 150, above the variance hinge in 30 of 30 seeds. The variance term becomes exactly zero by update 50, so covariance carries 100% of the loss after that point. On `arXiv`, the corresponding minimum is 0.43, below the hinge in 30 of 30 seeds. Thus, GV is variance-plus-covariance on `arXiv` and pure covariance after update 50 on `Products`; the cross-dataset analysis does not assume that the two losses are identical.

A.1.1 Training protocol and hyperparameter selection

All students use dropout 0.5 and Adam (Kingma & Ba, 2015) with $\beta_{\mathrm{Adam}} = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and no weight decay. The hint has its own trainable parameters, including the structural projector (and, for H2, the linear maps); we call this module the criterion and train it with a separate optimizer at learning rate 0.02. Student gradients are clipped at 20.0 and criterion gradients at 10.0. The student learning rate is 0.002 for both `arXiv` widths and 0.005 for `Products`.

The hint uses 8,192 sampled training nodes, a 16-dimensional structural embedding, temperature 0.075, and $\beta = 0.1$. Features are ℓ_2-normalized before structural projection. The matching teacher node is included among the contrastive candidates, and there is no student-to-student comparison. The teacher-free objective GV uses $\gamma_{\mathrm{var}} = \gamma_{\mathrm{cov}} = 1$, $\epsilon = 10^{-4}$ inside the square root, a variance target of 1, and raw (not normalized) student features on the same sampled nodes. The untrained twin CT is a GATv2 with the teacher's architecture, initialized with seed 777001 and never trained. The per-dimension hint H2 uses six linear maps from the 16-dimensional student layers to the teacher's 384-dimensional layers, trained at the criterion learning rate of 0.02. The permuted arms use one fixed node permutation per run, with seed 990000 plus the run seed. Output distillation uses $(\alpha, T) = (0.5, 4)$ with explicit division by the number of classes.

Width 16 inherits learning rate 0.002 from the earlier pipeline. Width 64, `Products` and `Mag` follow a frozen selection rule over $\{0.002, 0.005, 0.01\}$ using arm B only, three disjoint calibration seeds, the full 3,000-update schedule, mean fixed-final validation accuracy, and ties to the smaller rate. Test data are never consulted. Width 64 is selected by $+0.56$ pp. The `Products` choice is selected by $+0.019$ pp against per-candidate standard deviations of 0.09–0.19 at $n = 3$; this does not show that 0.005 is better than its neighbors. `Mag` selects 0.002. The pair $(\alpha, T) = (0.5, 4)$ is inherited from an earlier validation-only grid over $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$ and $T \in \{2, 4\}$ with three seeds and roughly 25 GPU-hours.

All main runs are full batch through a target-major `SparseTensor` adjacency. The graph is symmetrized and no self-loops are added. The auxiliary uses training nodes only. Cohorts use 30 paired seeds and complete-case analysis. Identity gates check the dataset, arm width, seed, runner hash, and teacher-cache hash. A failure removes the seed from every paired contrast.

Arm B runs the auxiliary forward pass with zero loss weight, so its structural-projector gradient is exactly zero at all 4,500 B-prefix updates. The Stage 5 preregistration text described B with the full loss; a dated clarification records that the executed zero-weight form governs.

A.1.2 Exposure-controller diagnostics

Raw-norm matching is not scale invariant across datasets. GV reads raw encoder features, while the contrastive term reads ℓ_2-normalized features. GV's matching coefficient is 0.16 on `arXiv` and 0.00027 on `Products`, a difference of about 600-fold. This difference concerns the unscaled

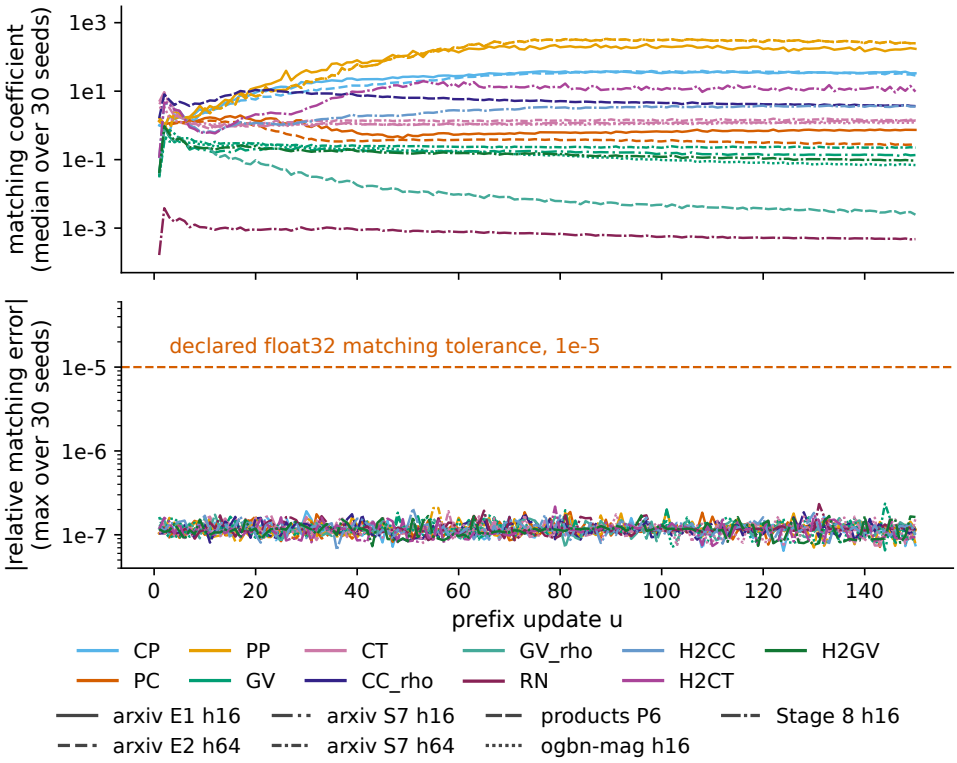


Figure 2: Exposure-matching diagnostics. The top panel shows the median matching coefficient across 30 seeds at each prefix update. The bottom panel shows the maximum relative matching error across the same seeds. All errors remain far below the declared float32 tolerance of 10^{-5}; coefficient size reflects the raw loss scale rather than the delivered exposure.

objective, not the exposure delivered. Across twelve matched arm–cohort pairs, coefficients range from 2.7×10^{-4} to 2.2×10^{2}, yet the maximum relative matching error is 2.4×10^{-7}, with zero clip events and zero guard flags. For dose matching, the real-teacher and teacher-free arms receive identical exposure to seven significant figures.

On `Mag`, the GV coefficient is 0.138 (median of per-seed medians; pooled mean 0.159), inside the preregistered range $[0.05, 0.50]$, and the CT coefficient is 1.12. The largest relative matching error is 1.75×10^{-7}, with no clip events and no guard flags across 120 runs.

A.2 Experimental setup

Datasets. We use the official OGB splits for `arXiv` and `Products` (Hu et al., 2020). The `arXiv` features are 128-dimensional averaged word2vec vectors from each paper's title and abstract. The graph is symmetrized.

The `Products` graph has 2,449,029 nodes and 123,718,152 undirected edge entries after symmetrization, which removes 128 duplicates. It has 47 classes and the 100-dimensional OGB features; the runner applies no PCA, normalization, or standardization. The split has 196,615 training nodes, 39,323 validation nodes, and 2,213,091 test nodes. For both datasets, the target-major `SparseTensor` adjacency is built after symmetrization and has no added self-loops.

The `Mag` cohort uses the homogeneous paper–citation projection of the graph: 736,389 paper nodes with 128-dimensional word2vec features (built as for `arXiv`) and 349 venue labels. The split is by year: training before 2018, validation 2018, and test 2019 and later (629,571 / 64,879 / 41,939 nodes). There are 5,416,271 citation edges, or 10,792,672 after symmetrization, with no isolated

nodes and no added self-loops. The venue label depends on node types that the projection drops (authors, institutions and fields of study), which is why every model has low accuracy on it.

Models and teacher checkpoints. Students are two-layer GraphSAGE models with width 16 or 64. Teachers are three-layer GATv2 models with 128 hidden units and three heads, or 384 concatenated features per layer. The `arXiv` teacher has 72.14% validation accuracy and 71.19% test accuracy. Its checkpoint records dropout 0.5, Adam at 0.005, no weight decay, 500 epochs, best epoch 497, and seed 0. The `Products` teacher has 91.40%, 90.42%, and 80.21% train, validation, and test accuracy. Its layer-wise full-neighborhood cache is built over training rows only. The `Mag` teacher is a GATv2 with the same architecture, trained with the `Products` recipe without changes (NeighborLoader, batch size 1024, fanout 10, dropout 0.5, learning rate 0.005, 150 epochs, best epoch 30 by validation). When evaluated one layer at a time using all neighbors, the teacher achieves 31.61% training accuracy, 33.37% validation accuracy, and 33.73% test accuracy. The teacher's cached representations for the training nodes and the untrained twin's cache (seed 777001) pass the same fidelity check used for the caches in the `oproducts` experiments . This 33.73% test accuracy is above OGB's homogeneous baselines on the same projection (GraphSAGE 31.53, GCN 30.43, MLP 26.92). We do not use the heterogeneous `Mag` benchmark on which (Ahluwalia et al., 2025) report. On `Mag`, the fixed sample of 8,192 nodes covers 1.30% of the training rows per update.

Main protocol. All main cohorts use 30 paired seeds and a fixed-final endpoint at update 3,000. The auxiliary is active for the first 150 updates. It is then removed, Adam is reset once in every arm, and training continues for 2,850 updates. Stage 6 E1 uses seeds 10000–10029, Stage 6 E2 uses 11000–11029, Stage 7 uses the same seeds in-fork, `Products` P3 uses 12000–12029, and its dose-matched P6 cohort uses 17000–17029, and `Mag` uses 19000—19029 (pilot seeds 19200–19202 are disjoint and enter no contrast). The noise, cross-entropy-only and per-dimension arms reuse 10000–10029 and 11000–11029 by design, so each new arm pairs within seed against the frozen Stage 6 references; the noise generator is seeded 800000 + seed. Calibration and smoke runs use disjoint reserved ranges. No seed was added, dropped, or rerun after examining any p-value.

All accuracy contrasts are paired within seed. We report two-sided Student-t 95% confidence intervals with $\mathrm{ddof} = 1$. The main GV/CT recovery fractions use 20,000 seed-keyed paired bootstrap resamples; the RN and H2 fractions use 10,000, and the `Mag` fractions use 20,000. We do not use an equivalence test. Historical experiments use $n = 5$ or $n = 3$ as indicated and a strict best-validation endpoint. They are clearly marked and are never pooled with the main cohorts.

A.2.1 Preregistration and analysis status

Table 6 separates confirmatory, descriptive, and historical evidence. Lock files record each preregistration's SHA-256 hash and timestamp. Analysis scripts recompute the hash and reject a mismatch or a run submitted before its lock. The descriptive analyses use records and checkpoint grids fixed by the original runners before these analyses existed. No preregistered gate, verdict, or claim depends on them.

`Mag` side-predictions. The Mag preregistration also made three side-predictions, reported here as they came out. The GV matching coefficient is inside the predicted range $[0.05, 0.50]$, and the geometry prediction held. The early-window prediction held in one part and failed in the other: the ratio of the CC advantage at removal to the final gain is 3.2 (predicted at least 3), but the advantage itself is $+2.65$ pp $[+2.02, +3.29]$, below the predicted $+3$ pp. Two small deviations from the preregistration affect no gate or outcome: the analyzer for the geometry prediction was fixed after the run, and the smoke run named for the GV coefficient was not done.

A.3 Audit of the routing comparison

We begin with a historical reproduction of the method that motivated the study. At its published settings, $\alpha = 0.9$ and $T = 4$, learned routing beats the logit-KD baseline by $+1.16$ pp at width 16 and $+0.42$ pp at width 64. At the same settings, plain cross-entropy with no distillation beats that baseline by 0.63 and 0.37 pp. Thus, the reported comparison uses a baseline that is worse than no distillation. This is an existence claim about one named configuration. The reproduction has $n = 3$, which is sufficient only for that check.

Table 6: Provenance of the result families. "Current" denotes the 30-seed mechanism pipeline. Historical rows use a different pipeline and endpoint.

Result	Status	n	Pipeline
Stage 6 factorial, both widths	Preregistered, locked analysis	30	current
Stage 4/5 correspondence controls	Preregistered	30	current
Stage 7 substitution and bounds	Preregistered, locked	30	current
Noise arm (RN)	Preregistered, locked outcome rule	30	current
Per-dimension hint (H2)	Preregistered, locked	30	current
Cross-entropy-only baseline (B0)	Preregistered, locked	30	current
`Mag` substitution	Preregistered, locked	30	current
`Products` P3 bound	Preregistered, locked gate (amended)	30	current
`Products` P6 dose-matched bound	Preregistered, locked, re-derived	30	current
Teacher probes	Preregistered verdict rule	3–5	historical
Real-versus-twin, six-layer teacher	Preregistered threshold	5	historical
Routing reproduction	Not preregistered	3	historical
Switch-off and decay analysis	Descriptive, post hoc	30	current, frozen records
Gap-closure normalization	Descriptive, post hoc	30	current, frozen records
Teacher-gap decomposition	Descriptive, post hoc	30	current, frozen records
Window, capacity, mini-batch, narrow `Products`	Historical, unlocked	5	historical
`snap-patents`, `roman-empire`	Historical, unlocked	5	historical

Table 7: Fixed-final test-accuracy effect of the real-teacher, correctly aligned hint. Values are paired mean differences in pp with 95% confidence intervals.

Cohort	CC − B	Positive seeds
Width 16 (E1)	$+1.0425$ $[+0.7662, +1.3189]$	28/30
Width 64 (E2)	$+0.3603$ $[+0.3018, +0.4188]$	29/30

With the corrected baseline, $\alpha = 0.5$, the six available run directories give 66.01 $\pm$ 0.03 for uniform all-to-all routing, 65.40 $\pm$ 0.13 for learned routing, and 64.92 $\pm$ 0.36 for the baseline at width 16. Learned routing is $+0.48$ pp above the baseline, while uniform routing is $+1.09$ pp above it. At width 64, every arm lies between 69.23 and 69.35. The corrected baseline minus plain cross-entropy is -0.140 $[-0.785, +0.505]$ at width 16 and -0.039 $[-0.676, +0.598]$ at width 64. At this budget, the corrected logit-distillation baseline cannot be distinguished from plain cross-entropy.

Uniform routing matches or exceeds learned routing in all six directories. The learned attention has collapsed to a one-hot pattern with entropy 1.5×10^{-6}. The audit student has baseline accuracy 64.9–66.0% at width 16, while the reported student has 70.87%. We therefore make no claim about the reported accuracy or about learned routing as a general mechanism. The matched-substitution instrument also cannot test learned routing because it requires constant, gradient-free attention.

A separate synthetic task with planted layer correspondence gives another reason for caution. Depth-2 and depth-3 groups are not identifiable by layer in 9 of 9 configurations. Probe accuracy rises monotonically toward the deepest teacher layer. For a six-layer, width-64, three-head MultiExitGAT teacher, attention heads exceed the second-best head by only $+0.003$ to $+0.011$ probe accuracy over three probe seeds $\{0, 1, 2\}$. The synthetic teacher is not the benchmark teacher, and $n = 3$.

A.4 Main effect on arXiv

The 2×2 factorial crosses the encoder target with the projector target. Table 7 reports the preregistered CC–B contrast. The uniform all-to-all hint improves over the already-distilled baseline at both widths. The effect is about three times larger for the narrower student.

A.5 Ablation I: schedule, reset, and node correspondence

The controls in Table 8 hold the optimization schedule fixed and change node assignment. Destroying correspondence costs about two pp, including when the permuted arm is gradient matched. A permuted hint is worse than no hint. The gain therefore cannot be explained by the reset, the extra forward pass or the mere presence of an auxiliary. Section 4.4.2 shows how much of it random noise explains. Correct node assignment matters.

Table 8: Correspondence controls. S is the correctly aligned hint. In $P_{\text{consistent}}$ the encoder and projector targets share one fixed node permutation; in P_{mismatch} they do not. Values are paired differences in pp with 95% confidence intervals.

Contrast	Estimate	Role
$S - B$ (Stage 4)	$+1.1056$ $[+0.7776, +1.4337]$, 29/30	Registered secondary
$S - B$ (Stage 5)	$+1.1967$ $[+0.9309, +1.4625]$, 30/30	Registered secondary
$S_{\text{perm}} - S$	-2.3975 $[-2.7039, -2.0912]$, 0/30	Registered primary
$S_{\text{perm_match}} - S$	-1.9853 $[-2.3257, -1.6449]$, 0/30	Registered secondary
$P_{\text{consistent}} - P_{\text{mismatch}}$	$+1.8679$ $[+1.3960, +2.3399]$, 29/30	Registered secondary
$P_{\text{consistent}} - S$	-0.2546 $[-0.4470, -0.0623]$, 11/30	Registered primary

Table 9: Factorial interaction and simple effects. Values are paired differences in pp with 95% confidence intervals.

Quantity	Width 16	Width 64
Interaction	$+1.6592$ $[+1.3153, +2.0030]$	$+0.5845$ $[+0.4696, +0.6994]$
d_{permuted}	$+1.6205$ $[+1.3224, +1.9186]$, 30/30	$+0.5932$ $[+0.4998, +0.6866]$, 30/30
d_{correct}	-0.0387 $[-0.1563, +0.0789]$, 15/30	$+0.0086$ $[-0.0557, +0.0730]$, 18/30
PC $-$ CC	-2.0152 $[-2.3791, -1.6513]$, 0/30	-0.7024 $[-0.7898, -0.6149]$, 0/30

The factorial localizes this harm. Table 9 shows a large interaction, but it is carried entirely by the permuted-encoder branch. When the encoder target is correct, permuting the projector target changes nothing at either width. The encoder side is load-bearing; the projector side is not. The CP–CC contrast also provides a preregistered internal null.

Figure 4 (left) visualizes the factorial simple effects and interaction at both widths.

A.6 Ablation II: gradient-matched source substitution

Section 4.4.2 and Table 3 report the substitution results; this section adds the seed pairing and the tuning budget behind them.

The bounds on what the real teacher adds beyond GV, $+0.13$ pp at width 16 and $+0.07$ pp at width 64, are about one-eighth and less than one-fifth of the effects they are used to explain.

This result does not claim equivalence. No equivalence margin was registered. It shows that the data do not distinguish the substitutes from CC within the stated bounds. The width-16 GV and CT arms use seeds 10000–10029 and are paired with Stage 6 E1 B and CC. At width 64, all arms come from the Stage 7 fork on seeds 11000–11029.

The tuning budget favors the baseline. Its (α, T) pair comes from an eight-configuration, three-seed validation search. GV has no objective-specific tuning; its embedding dimension and epsilon are fixed a priori, and its scale is set only by the exposure controller. This asymmetry would limit a negative conclusion about GV. It does not weaken the observed recovery because both untuned substitutes succeed.

A.6.1 Additional noise and per-dimension controls

RN replaces the structured target with isotropic Gaussian noise in the encoder gradient for the same 150 updates. It uses a private generator and is rescaled at every update to the same-seed CC norm. RN recovers 0.765 $[0.634, 0.897]$ of the gain: $+0.7973$ pp $[+0.5413, +1.0533]$ over B, and -0.2453 pp $[-0.4002, -0.0903]$ relative to CC. Thus, perturbation alone recovers much of the effect, but a structured target that need not be trained accounts for the remaining gap. Before the RN runs, its preregistration named two possible outcomes, a recovery fraction above 0.9 or below 0.5, and predicted the second. The observed fraction, 0.765 [0.634, 0.897], lies between the two and excludes both, so neither predicted outcome occurred: noise of the same size reproduces most of the gain, but not all of it.

Table 10: Within-window test-accuracy advantage over B and fixed-final advantage, in pp. Each entry is a paired mean over 30 seeds. The auxiliary is active through update 150 and absent afterward.

Cohort	Arm	$u = 1$	25	50	75	100	125	150	Final
arXiv, width 16	CC	$+0.05$	$+0.28$	$+1.11$	$+2.76$	$+5.57$	$+5.98$	$+5.66$	$+1.043$
arXiv, width 16	CP	$+0.05$	-1.08	-2.25	-1.43	-0.53	$+0.66$	$+1.36$	$+1.004$
arXiv, width 16	PC	$+0.03$	-2.51	-4.89	-4.62	-4.79	-4.25	-3.01	-0.973
arXiv, width 16	PP	$+0.02$	-0.83	-2.96	-2.99	-2.82	-2.38	-1.16	$+0.648$
arXiv, width 64	CC	-0.02	$+0.19$	$+5.02$	$+3.90$	$+2.95$	$+2.12$	$+1.87$	$+0.353$
arXiv, width 64	CT	-0.02	$+0.68$	$+5.93$	$+3.45$	$+2.41$	$+1.60$	$+1.35$	$+0.323$
arXiv, width 64	GV	-0.02	$+0.47$	$+3.95$	$+2.36$	$+1.95$	$+1.38$	$+1.20$	$+0.353$
arXiv, width 64	PP	$+0.01$	-0.61	$+0.27$	$+0.47$	$+0.46$	$+0.28$	$+0.22$	$+0.236$
Products	CC	-0.00	$+0.39$	$+0.41$	$+0.32$	$+0.24$	$+0.22$	$+0.19$	$+0.015$
Products	GV	$+0.00$	-0.01	$+0.11$	$+0.10$	$+0.06$	$+0.07$	$+0.08$	-0.002
arXiv, width 16	RN	-0.02	-2.17	-6.87	-4.72	-5.32	-4.55	-3.37	$+0.797$
arXiv, width 16	H2CC	$+0.05$	$+2.33$	$+2.89$	$+4.61$	$+6.86$	$+6.75$	$+6.19$	$+1.070$
arXiv, width 16	H2CT	$+0.05$	$+2.03$	$+1.60$	$+2.33$	$+4.60$	$+4.70$	$+4.34$	$+1.051$
arXiv, width 16	H2GV	$+0.02$	$+0.86$	$+1.66$	$+2.56$	$+3.99$	$+4.22$	$+3.96$	$+0.989$
Mag	CC	-0.02	$+0.43$	$+0.96$	$+1.43$	$+2.04$	$+2.97$	$+2.65$	$+0.828$
Mag	GV	-0.00	$+0.69$	$+1.06$	$+0.57$	$+0.66$	$+0.98$	$+0.88$	$+0.726$
Mag	CT	-0.01	$+0.22$	$+0.98$	$+1.22$	$+2.03$	$+2.93$	$+2.61$	$+0.787$

Table 11: Switch-off and fixed-final advantages over B, in pp. Each value is a paired mean over 30 seeds. Spread is the largest minus the smallest value in the row.

Cohort and arms	Switch-off values	Spread	Final values	Spread
arXiv, width 16, CC and CP	$+5.662, +1.357$	4.31	$+1.043, +1.004$	0.039
arXiv, width 64, CC/CT/GV	$+1.874, +1.353, +1.202$	0.67	$+0.353, +0.323, +0.353$	0.030
arXiv, width 16, CC and RN	$+5.662, -3.365$	9.03	$+1.043, +0.797$	0.246
Mag, width 16, CC/CT/GV	$+2.655, +2.610, +0.880$	1.78	$+0.828, +0.787, +0.726$	0.102

The main contrastive hint reduces each node representation to a scalar mean. We therefore also test a per-dimension family. For each of the six teacher–student layer pairs, a learned linear regressor maps the student hidden state to the teacher's 384-dimensional state, and mean-squared error is averaged over pairs. H2CC receives the contrastive CC exposure through the ρ controller; H2CT and H2GV match H2CC's trace. Their paired mean advantages over B at switch-off are $+6.19$, $+4.34$, and $+3.96$ pp, and their fixed-final advantages are $+1.070$, $+1.051$, and $+0.989$ pp, respectively. Their recovery fractions are 0.98 $[0.89, 1.08]$ for H2CT and 0.92 $[0.80, 1.04]$ for H2GV. Both point estimates are above the preregistered 0.9, although both intervals extend below it. Figure 1b shows their intervals, and Table 10 gives their trajectories.

A.7 Switch-off analysis: the early advantage does not consistently rank the final effects

The confirmatory endpoint is update 3,000, but the auxiliary stops at update 150. Table 10 reports test-accuracy differences from B at the checkpoint grid written by the original runners. An independent reevaluation of 4,800 frozen checkpoints uses 0.205 GPU-hours and reproduces 3,597 of 3,600 gated metric comparisons exactly. The analysis is descriptive and post hoc. The checkpoint grid was fixed before this analysis. The Products GV arm matches CC's applied encoder-gradient norm to relative error 1.9×10^{-7}; its small coefficient measures the unscaled objective, not the delivered exposure.

The real-teacher width-16 advantage is $+5.66$ pp at switch-off and $+1.043$ pp at the final endpoint. Only 18.4% remains. At width 64, $+1.87$ pp falls to $+0.353$ pp, or 18.8%. The same broad decay appears for every correctly aligned target. The size of the switch-off advantage does not consistently order the endpoints.

At width 16, CC and CP are 4.31 pp apart at switch-off and 0.039 pp apart at the endpoint (Table 11). At width 64, CC, CT, and GV span 0.67 pp at switch-off and only 0.030 pp at the endpoint. PC and PP do not follow this convergence because their encoder target is permuted; their final effects are -0.973 and $+0.648$ pp. This agrees with the preregistered interaction.

Table 12: Baseline training progress as a percentage of its own fixed-final training accuracy.

Cohort	$u = 1$	25	50	75	100	125	150
arXiv, width 16	5.0%	31.6%	42.1%	51.0%	60.1%	67.0%	72.0%
arXiv, width 64	9.4%	37.4%	54.4%	68.3%	75.8%	80.4%	82.8%
Products	5.8%	68.3%	85.5%	90.7%	93.2%	94.6%	95.4%
Mag	2.7%	16.8%	35.8%	43.9%	50.4%	56.1%	61.8%

Table 13: Paired endpoint differences on training, validation, and test accuracy, in pp.

Cohort	Δtrain	Δvalid	Δtest
arXiv, width 16, CC − B	$+1.410$ [$+1.044, +1.775$]	$+1.096$ [$+0.819, +1.372$]	$+1.043$ [$+0.766, +1.319$]
arXiv, width 64, CC − B	$+0.891$ [$+0.794, +0.989$]	$+0.320$ [$+0.251, +0.389$]	$+0.360$ [$+0.302, +0.419$]
Products, CC − B	-0.007 [$-0.032, +0.018$]	-0.037 [$-0.081, +0.007$]	$+0.015$ [$-0.129, +0.159$]
Products at arXiv dose, CC$_\rho$ − B'	$+0.004$ [$-0.019, +0.028$]	$+0.016$ [$-0.040, +0.072$]	$+0.129$ [$-0.033, +0.290$]
Mag, CC − B	$+0.968$ [$+0.691, +1.244$]	$+0.936$ [$+0.675, +1.197$]	$+0.828$ [$+0.588, +1.068$]

Three simple summaries are false. First, the final effect is not always a fixed fraction of the switch-off effect. Across arms, the ratio ranges from -55.9% for width-16 PP, which changes sign, to $+105.5\%$ for width-64 PP. The 18% ratio is a property of CC, not a universal constant. Second, the peak has no fixed location. Width 64 and Products peak at update 50, while width 16 peaks at update 125. Third, peak height does not order endpoints. At width 64, CT peaks at $+5.93$ pp and finishes below CC, while GV peaks at $+3.95$ pp and finishes level with CC.

A.7.1 How far the baseline has trained by switch-off

Table 12 reports baseline training accuracy at each checkpoint as a percentage of its own final training accuracy. Products is already at 68.3% by update 25 and 95.4% by update 150. Its transient is significant at update 25 ($+0.392$ pp) and update 50 ($+0.409$ pp), but it is about fifteen times smaller than the width-16 arXiv transient. Mag is the least trained of the four cohorts at switch-off (61.8% of its final training accuracy, the same as its pilot estimate). This is the property that the preregistered rule used to predict a gain on Mag.

A.7.2 The endpoint gain is a fitting gain

At the endpoint, the hint raises training accuracy more than test accuracy on arXiv. This is not the pattern expected from a simple variance-reducing regularizer. Table 13 gives the exact decomposition.

At width 16, the seed-wise change in training accuracy and change in test accuracy have Spearman correlation $+0.786$ ($p < 0.0001$). The correlations are $+0.287$ ($p = 0.124$) at width 64 and $+0.192$ ($p = 0.308$) on dose-matched Products. The two preregistered internal nulls give $+0.111$ ($p = 0.561$) and $+0.245$ ($p = 0.192$). Both variables are paired differences. We do not correlate a contrast with its baseline level because that statistic is biased by construction. Under exchangeable arms at $n = 30$, its expected Spearman correlation is -0.672 with a 95% range $[-0.849, -0.418]$. Two known-null contrasts in our records give -0.669 and -0.671.

A.8 Historical checks: window placement and capacity

These experiments use $n = 5$, an earlier pipeline, and a best-validation endpoint. They support parts of the mechanism account but are not pooled with the main cohorts.

Window placement. Extending an early window from 150 to 300 or 750 updates does not increase the endpoint effect. This matches the later trajectory, which plateaus near update 125. However, a late window still produces a small gain. That result is not explained by the trajectory analysis and remains open.

The window from updates 1–30 gives $+0.644$ pp because it ends before the later trajectory reaches its plateau. The 1–150, 1–300, and 1–750 windows give $+0.806$, $+0.789$, and $+0.783$ pp. A

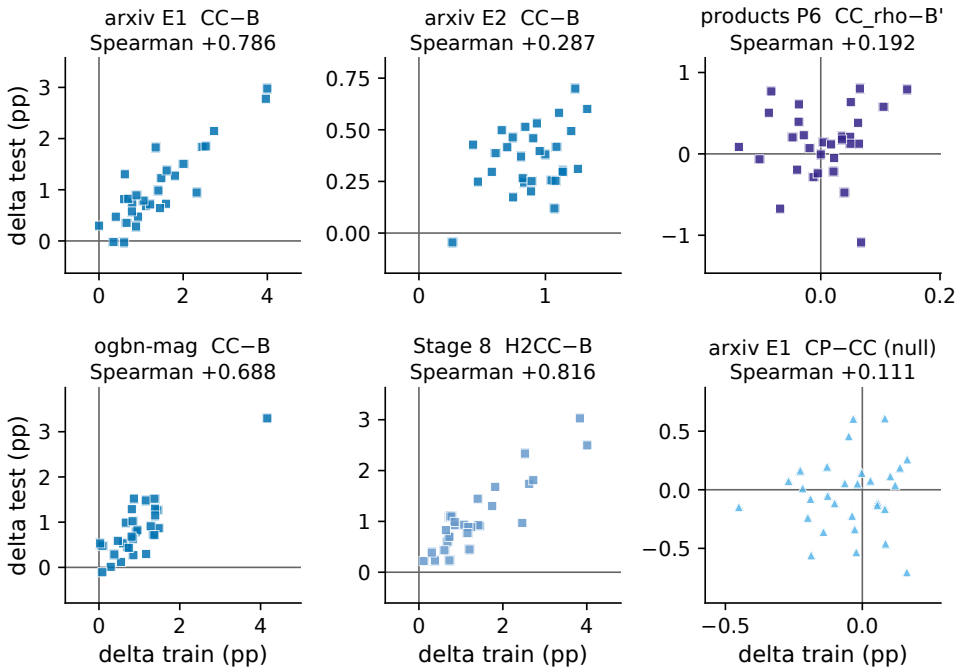


Figure 3: Seed-wise endpoint changes in training and test accuracy. The Spearman correlations are $+0.786$ for `arXiv` width 16, $+0.287$ for `arXiv` width 64, $+0.192$ for dose-matched `Products`, $+0.688$ for `Mag`, $+0.816$ for the per-dimension hint H2CC, and $+0.111$ for the preregistered CP–CC null. Both axes are paired within-seed differences.

Table 14: Historical window-placement effects in pp, each with $n = 5$.

Window	Estimate	Positive
Updates 1–30	$+0.644$ $[-0.021, +1.309]$	5/5
Updates 1–150	$+0.806$ $[+0.020, +1.592]$	5/5
Updates 1–300	$+0.789$ $[+0.120, +1.458]$	5/5
Updates 1–750	$+0.783$ $[+0.150, +1.415]$	5/5
Updates 150–300 (delayed)	$+0.601$ $[+0.041, +1.160]$	5/5
Last 150 updates	$+0.222$ $[+0.046, +0.398]$	5/5
Early 150 − delayed 150	$+0.205$ $[-0.242, +0.652]$	4/5

delayed 150-update window gives $+0.601$ pp. A window over the last 150 updates gives $+0.222$ pp $[+0.046, +0.398]$, 5/5, even though the student is nearly converged. This last result is 3.6 times smaller than the early 150-update result, but it is not explained by the main trajectory analysis.

Capacity ladder. The effect falls as the student widens from 16 to 64 and then remains near $+0.2$ pp at widths 128–512. The width-512 interval spans zero.

Figure 4 (right) summarizes the historical window-placement and capacity results.

Other historical controls. Under a neighbor-sampled mini-batch pipeline on `arXiv`, the hint gives $+2.544$ pp $[+0.083, +5.005]$, 5/5. Thus, mini-batching does not explain the `Products` bound; both main mechanism cohorts are full batch in any case. On a width-8 `Products` student, forcing exposure ratio 0.65 gives -1.583 pp $[-2.251, -0.915]$, 0/5. This run uses a different pipeline, a sustained hint, and evaluates the checkpoint with the best validation performance. It provides the only evidence in the record for a width-8 student on `Products`.

Earlier width-16 source controls on `arXiv` give the following effects:

Table 15: Historical capacity ladder, each with $n = 5$.

Student width	Estimate	Positive
128	$+0.2226$ $[+0.0430, +0.4022]$	4/5
256	$+0.1333$ $[+0.0484, +0.2183]$	5/5
512	$+0.2469$ $[-0.3095, +0.8033]$	3/5

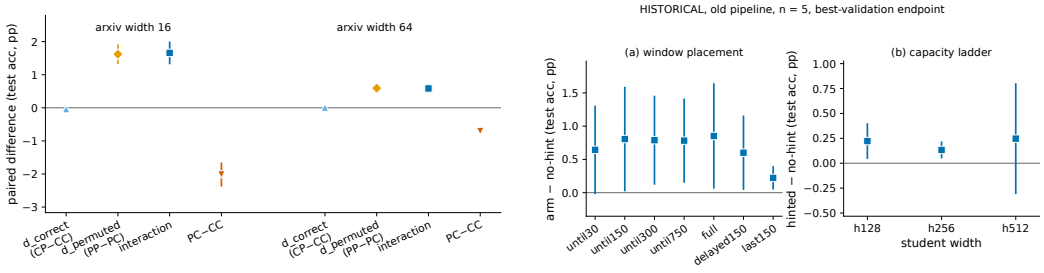


Figure 4: Additional ablation results shown side by side. Left: factorial simple effects and interaction at widths 16 and 64, with paired 95% confidence intervals. Right: historical window-placement and capacity-ladder effects, each with $n = 5$ and a best-validation endpoint.

- real-teacher contrastive hint: $+0.83$ $[+0.03, +1.63]$;
- untrained twin: $+0.55$ $[+0.09, +1.00]$;
- teacher-free variance–covariance: $+0.67$ $[+0.12, +1.22]$, 5/5;
- real-teacher frozen per-dimension MSE: $+0.52$ $[-0.19, +1.22]$;
- frozen random input projection with a contrastive loss: $+0.03$ $[-0.21, +0.27]$;
- the same random projection with MSE: $+0.33$ $[-0.52, +1.17]$.

Correct node correspondence alone is not sufficient: the random projection has correspondence but no graph propagation and does not recover the gain. This is the only evidence here that the target may need graph structure, and it has $n = 5$ on a different pipeline.

A.9 Main cross-dataset bound on Products

Everything is inherited from the `arXiv` pipeline except the student learning rate, which follows the frozen selection rule in Appendix A.1. At natural exposure, the registered CC–B gate does not pass. We therefore report bounds and do not compute a recovery fraction.

The P3 preregistration anticipated that raw-norm matching would give GV a small coefficient: 0.021 in the smoke run and a cohort median 2.8×10^{-4}, compared with 1.6×10^{-1} on `arXiv`. The coefficient measures the unscaled size of GV, not its delivered gradient. `Products` encoder features have roughly twenty times the variance of `arXiv` features. GV uses raw features, while the contrastive objective uses normalized features. The same controller therefore produces a coefficient about 600 times smaller.

The applied intervention is matched exactly. GV and CC both have integrated applied norm 1.3700. The median relative error is zero, the maximum is 1.9×10^{-7}, and no guard flag is raised over 30×150 updates.

The same matching convention scales the permuted `arXiv` arms upward by factors of 135 and 221. Those arms are not treated as over-driven because their large coefficients, like GV's small coefficient, only compensate for the unscaled gradient. In every case, the applied norm matches the reference.

GV is also not inert. At update 150, relative to B, its participation ratio changes by $+1.304$, compared with $+0.665$ for CC. Its mean variance changes by -1.401, compared with -0.819; mean ℓ_2 norm by -0.572, compared with -0.329; mean absolute off-diagonal correlation by -0.0289, compared with -0.0169; and mean pairwise cosine by $+0.0099$, compared with -0.0018. Every quantity is

Table 16: Natural-exposure results on Products. Values are fixed-final paired differences in pp with 95% confidence intervals. A $\leq$ entry is the upper limit of the 95% confidence interval.

Contrast	Estimate	Verdict
CC − B (gate family)	$+0.0147$ $[-0.1292, +0.1586]$, 14/30	Gate not passed; CC − B $\leq +0.16$ pp
GV − B (registered primary)	-0.0024 $[-0.0788, +0.0740]$, 16/30	$\leq +0.07$ pp; exposure matched exactly
GV − CC	-0.0171 $[-0.1526, +0.1183]$	Bound only
Fraction	NOT COMPUTED	CC − B did not pass

Table 17: Dose-matched results on Products. The frozen arXiv exposure ratio is imposed per update. A $\leq$ entry is the upper limit of the 95% confidence interval.

Contrast	Estimate	Verdict
$CC_\rho - B'$	$+0.1288$ $[-0.0326, +0.2901]$, 20/30	$\leq +0.29$ pp
$GV_\rho - B'$	-0.0362 $[-0.1640, +0.0916]$, 15/30	$\leq +0.09$ pp
$GV_\rho - CC_\rho$	-0.1650 $[-0.3315, +0.0015]$, $p = 0.052$	Bound only

significant at $n = 30$. GV moves the measured geometry more than CC but changes accuracy by only -0.0024 pp $[-0.0788, +0.0740]$.

At the final checkpoint, every geometry contrast is null for both GV and CC. The 2,850 auxiliary-free updates erase the measured geometric signature while an accuracy residue can remain. Thus, the final effect is not carried by a persistent change in these geometry measures.

The Products has no untrained-twin or permuted arm. The teacher-free arm addresses source-free substitution under both exposure conventions, but the untrained-twin substitution remains untested on Products.

A.10 DOSE-MATCHED PRODUCTS

The natural Products exposure ratio is 0.220, compared with 0.736 on arXiv; it is 3.3 times lower. We therefore impose the frozen arXiv width-16 exposure trajectory update by update. The delivered integrated encoder norm is 7.666, compared with arXiv's 7.468, a difference of $+2.7\%$. The controller matches the target ratio at 9,000 of 9,000 run–update points, with maximum recomputed relative error 1.86×10^{-7}. This requires roughly fivefold scaling of the real-teacher gradient.

The achieved median ratio is 0.7414148; the driving trace has median 0.7414147. CC_ρ and GV_ρ match exposure to seven significant figures. Imposing arXiv's exposure yields $+0.13$ pp $[-0.03, +0.29]$, not the $+1.04$ pp arXiv effect. This rules out under-dosing as the full explanation for the cross-dataset difference, but it does not exclude a small residual effect. The upper bound is 80% of the width-64 arXiv effect, so the difference is one of effect size, not a proven presence-versus-absence split. Under the same intervention, GV_ρ raises the 16-dimensional participation ratio by $+4.31$ and changes training accuracy by only $+0.0044$ pp.

A.11 DOES THE TEACHER CONTAIN EXTRA INTERMEDIATE INFORMATION?

Frozen-teacher probes fit identical heads to logits, one hidden layer, their concatenation, and a row-shuffled control. A rule fixed in advance declares incremental information only if the concatenation improves validation negative log-likelihood over both the logits and the shuffled control.

For the three-layer arXiv teacher, the largest apparent gain is at the empirical floor and is explained by one hidden layer being an invertible relabeling of another. We report a bound of ≤ 0.01 nats and ≤ 0.3 pp. A graph-aware probe gives the same result, and one node-degree feature added to the logits beats every hidden layer. The Products controls move by only $+0.034$ to $+0.055$, and the probe is null.

The six-layer teacher is the exception. Its head is over-smoothed, and concatenating a hidden layer with the logits changes validation negative log-likelihood by -0.058 nats and accuracy by $+1.46$ pp, beating every control. A follow-up compares the real-teacher hint with an untrained twin.

The difference is only $+0.16$ pp, below the preregistered no-transfer threshold of 0.3, at $n = 5$. This historical result uses uniform all-to-all routing rather than a layer-targeted hint. It shows that probe-detectable information can exist without being transferred by the tested hint.

A.12 Why the Products effect is small

Section 4.5 gives two reasons why little acceleration is available on Products; this section gives the numbers behind them. By update 150, Products has reached 95.4% of its final training accuracy (Table 12), and its switch-off advantage of $+0.19$ pp is already past its peak (Table 10). The dose-matched estimate, $+0.13$ pp $[-0.03, +0.29]$, fits this account: the upper confidence limit on any remaining `Products` gain is $+0.29$ pp. On the training rows the hint sees, the Products teacher leads the student by only 0.85 pp, against 8.94 pp on test (Table 2); on arXiv the validation and test gaps are 6.14 and 6.01 pp.

The OGB `Products` split is ordered by sales rank: the top 8% of nodes form training, the next 2% validation, and the remaining 90% test (Hu et al., 2020). Under a random split with the same ratio, GraphSAGE reaches 88.20% test accuracy with about one pp of generalization gap, compared with 78.50% and roughly fourteen pp under sales-rank ordering. OGB reports valid-to-test gaps of 14.48 pp for an MLP, 17.83 for Node2vec, 16.36 for GCN, and 13.74 for GraphSAGE. The student here has a 17.94 pp gap. Of its 8.94 pp teacher–student test gap, 7.73 pp is excess distribution-shift damage beyond the teacher's own valid-to-test drop. A hint computed on training rows cannot directly observe this gap.

`Products` is also strongly limited by input representation. Correct & Smooth reaches 84.18% with an MLP and 137 times fewer parameters than the then-best GNN (Huang et al., 2021). GIANT features move an MLP from 61.06% to 79.82%, above a full GraphSAGE (Chien et al., 2022). Label propagation without trainable parameters reaches 74.34%, three pp above this student.

GLNN, a GNN-to-MLP method, shows the same asymmetry in three papers: it gains $+6.39$ pp on `Products` and $+7.41$ pp on `arXiv` in its own paper (Zhang et al., 2022), $+3.69$ and $+8.08$ pp in the reproduction reported by VQGraph (Yang et al., 2024), and $+3.29$ and $+13.13$ pp in the one reported by HGMD (Wu et al., 2024). GraphAKD is a counterexample: it reports $+6.46$ pp on `Products` and $+1.62$ pp on `arXiv` (He et al., 2022). Its `Products` student starts 9.60 pp behind its teacher, while its `arXiv` student starts 1.31 pp behind. This is consistent with a larger effect when more fit is available, but it is not a controlled test of our mechanism.

A.13 Negative results and scope limits

`snap-patents`. The natural hint harms accuracy by -2.60 pp $[-2.86, -2.33]$. The untrained twin harms by the same amount: twin minus natural is -0.10 pp $[-0.35, +0.15]$. Clamping the exposure ratio does not rescue it. This is a historical $n = 5$ mini-batch run with a sustained hint and a teacher weaker than its student, 31.01% versus 37.47%. The graph is strongly heterophilous, with edge homophily 0.073. We found no published distillation result on this dataset. The result is a first data point, not a confirmation.

`roman-empire`. The effect is $+0.10$ pp $[-0.05, +0.26]$ at $n = 5$. The teacher has 57.66% accuracy, about twenty pp below its student at roughly 77%. Increasing exposure is harmful: ratio 1.5 gives -0.33 pp, ratio 3 gives -1.75 pp, and ratio 6 gives -4.22 pp. An earlier version is void because of a teacher-loading defect and split leakage. The valid run uses Platonov split column zero (Platonov et al., 2023); `snap-patents` uses LINKX official split zero (Lim et al., 2021).

A.14 Reproducibility and integrity checks

We re-verify 13,382 checkpoint hashes against four package inventories and runner write-time records: 13,382 match, zero mismatch, and zero are absent. Independent reevaluation covers 600 runs at update 150 and final checkpoints on train, validation, and test, for 3,600 comparisons. Of these, 3,597 match exactly. The three differences are single-node argmax flips in width-64 `arXiv` runs caused by non-associative scatter; the largest is 2.056×10^{-3} pp. All 540 `Products` comparisons match exactly. This numerical floor is more than two orders of magnitude below the main `arXiv` effects ($+0.36$ and $+1.04$ pp).

Every main cohort is complete at 30 seeds. The remaining cohorts were rerun cleanly, and both logs are retained. Three package files fail a naive checksum by design: a manifest cannot hash itself, and two reports are written after the manifest is sealed. Three prose descriptions were corrected after source inspection: one coefficient did not measure what its prose claimed, one experimental condition on `products` had been called uninformative despite exactly matched exposure, and attention-path parameters in the optimizer receive no gradient. None changes a reported number.

The factorial implementation has 22 positive invariants and five must-fail fixtures. The checks cover the permutation, sampling closure, shared routed gradients, exact norm matching, unit CC scale, and the scope of projector replacement. They also cover the zero-target policy, optimizer rebinding after reset, and the zero structural-projector gradient for B while its auxiliary forward pass runs. The five failing fixtures inject a shifted positive index, a wrong projector group, a double replacement, a one-byte reference change, and a 5% achieved-norm alteration. All 27 checks pass. Analyzer suites add 27 Stage 7 checks, 34 Stage 6 checks, and 27 checks for the analysis of experiments on `products`.

Compute. Each job uses one NVIDIA A100. Scheduler accounting covers 1,765 job records from 1–19 September 2026. Stage 4 uses 8.52 GPU-hours and Stage 5 uses 8.95. Stage 6 E1 uses 11.13 plus 0.04 calibration; E2 uses 12.60 plus 0.04. Stage 7 uses 4.77 for width 16 (60 runs including smokes and checks) and 12.99 for width 64 (150 runs). `Products` P3 uses 35.74 and P6 uses 35.93. The noise, cross-entropy-only and per-dimension arms use 13.58 (six arrays of 30 runs). `Mag` uses 29.75 for the cohort (120 runs, about 0.25 GPU-hours each), plus 2.37 for learning-rate calibration and 0.72 for the three pilot runs. As for the other datasets, teacher training and cache construction (8.56) are not included. Total main-pipeline compute is 177.13 GPU-hours. This excludes the earlier pipeline, teacher training, and teacher/twin cache construction.

Environment. Runs use Python 3.10.20, PyTorch 2.7.1+cu118, CUDA 11.8, cuDNN 9.1.0, PyTorch Geometric 2.7.0, OGB 1.3.6, and NumPy 2.2.6. Hardware is NVIDIA A100 40GB and 80GB PCIe with driver 595.71.05. `cudnn.deterministic` is enabled and benchmarking is disabled. Scatter and atomic reductions remain non-associative in float32, which causes the disclosed reevaluation floor. Bitwise reproduction across environments is unavailable. The in-fork versus frozen diagnostic finds no systematic offset, and no preregistered contrast crosses environments.

B Related Work

Knowledge distillation and its mechanism. Knowledge distillation trains a smaller student using signals from a larger teacher. Output-level methods use the teacher's predictions (Hinton et al., 2014), while intermediate methods also align hidden representations (Romero et al., 2015; Tian et al., 2020). An improvement from an intermediate objective, however, does not by itself show that useful teacher representations were transferred. The additional loss also changes the student's gradients and optimization path. Prior studies show that distillation can help with randomly initialized or weak teachers and that a larger teacher can sometimes harm a smaller student (Sarnthein et al., 2023; Yuan et al., 2020; Cho & Hariharan, 2019). These findings question a purely information-transfer explanation, but they do not separate teacher information from the optimization effect of a specific intermediate objective.

DETAIL addresses this gap. It does not introduce a new distillation objective. Instead, it replaces the source of an existing intermediate target while matching the measured encoder-gradient norm at every update. This intervention tests whether the gain requires learned teacher information or can arise from source-independent optimization pressure.

Intermediate GNN distillation. Intermediate GNN objectives may transfer node embeddings, local relations, or global graph structure (Tian et al., 2025). Local Structure Preserving distillation matches relations among neighboring nodes (Yang et al., 2020). G-CRD uses contrastive learning to preserve global topology in the teacher's embedding space (Joshi et al., 2024). Geometric Knowledge Distillation aligns topology across teacher and student representations (Yang et al., 2022). GraphAKD aligns teacher and student node-level and class-level representations through adversarial training (He et al., 2022). These methods study different forms of transferable knowledge, but they do not control the strength of the auxiliary gradient while changing its information source. STRIDE

combines embedding and structural losses across teacher and student layers, using learned attention to weight each layer pair (Ahluwalia et al., 2025). Its central idea is that intermediate layers capture complementary information about node representations and graph structure. Combining these signals aims to provide richer supervision for a compact GNN.

GNN to MLP distillation. A separate line of work distills graph knowledge into multilayer perceptrons for graph-free inference. GLNN distills a GNN directly into an MLP (Zhang et al., 2022). NOSMOG studies the accuracy, robustness, and efficiency of MLPs trained with graph-derived signals (Tian et al., 2023). VQGraph uses a vector-quantized representation space to connect GNN and MLP representations (Yang et al., 2024). HGMD studies settings in which a harder teacher produces a poorer student (Wu et al., 2024).

These studies provide useful cross-dataset evidence, but they do not offer controlled comparisons between `arXiv` and `Products`. GLNN's gains on `arXiv` and `Products` are $+7.41$ and $+6.39$ pp in its own paper, $+8.08$ and $+3.69$ pp in the reproduction reported by VQGraph, and $+13.13$ and $+3.29$ pp in the one reported by HGMD. GraphAKD provides a different pattern, with gains of $+1.62$ pp on `arXiv` and $+6.46$ pp on `Products`. Its starting teacher and student gaps are 1.31 and 9.60 pp, respectively. These values cannot identify a mechanism because the methods, model architectures, baselines, and evaluation endpoints differ.