PATHWAY-ALIGNED SUPERVISION: ASSIGNING NEGATIVE SAMPLERS IN TEMPORAL LINK PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal link predictors combine node context with pair history, yet their scoring pathways usually learn from the same negative examples. A comparison that distinguishes node representations can provide no direct gradient to an independently parameterized, deterministic history residual: identical history inputs cancel that residual in a pairwise score difference. This mismatch motivates Pathway-Aligned Supervision (PAS), which assigns random negatives to the node-score loss and historical negatives, with random fallback, to the full pair-score loss. With the scoring architecture fixed, we compare all four assignments and a validation-selected shared mixture. On tgbl-review, PAS raises the test mean reciprocal rank of a TGAT-based predictor from 0.20 under all-random supervision to 0.49 and exceeds the selected mixture; disjoint seeds confirm the gain. Better training contrast does not guarantee better ranking, however: PAS hurts tgbl-wiki, and changing evaluation candidates can reverse its advantage without retraining. Component experiments reveal a further distinction: a fitted history residual can help ranking even when changing its negative supervision harms the complete predictor. These findings support assigning negatives at the component level while evaluating the assignment under the intended ranking task.

1 INTRODUCTION

Temporal link prediction uses past interactions to rank the destination of a future event. Useful evidence exists at more than one level. A destination may be broadly active, or it may have a specific history with the query source. For example, an editor may visit a popular page for the first time or return to a page they have edited before. Temporal encoders describe the participating nodes, while recurrence-aware features describe the pair (Xu et al., 2020; Rossi et al., 2020; Poursafaei et al., 2022; Yu et al., 2023; Lu et al., 2024). Neither source alone describes every interaction. Combining them also raises a training question: whether the same comparisons teach both components what they need to learn.

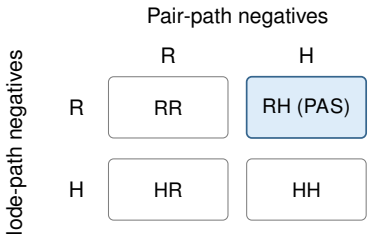
A predictor learns to use this evidence through comparisons with negative destinations. Random negatives provide comparisons across the destination population; historical negatives select previous partners other than the observed target. Benchmark studies use historical candidates to challenge predictors beyond easy random comparisons (Poursafaei et al., 2022; Huang et al., 2023; Gastinger et al., 2024), while studies of harder training negatives report both gains and losses (Yu et al., 2023; Gupta & Bedathur, 2024; Chen et al., 2024; Yi et al., 2025a). Choosing one sampler, or one shared mixture, asks which comparisons suit the whole predictor. It leaves open whether its scoring components need the same negative evidence.

The same comparison can teach different parts of a predictor different things. Consider a score combining node context with an additive term for pair count and recency. If the positive and negative candidates have both never interacted with the source, their history inputs are identical. The independent history terms cancel in a score difference, although the node contexts may remain distinguishable. A previous partner supplies a history contrast that this comparison lacks. Such a contrast can matter even when the target is new: the target may still have to outrank familiar alternatives. This motivates studying which negatives train each score component, rather than treating pair recurrence alone as a sampler-selection rule.

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.

We examine this question with a fixed two-pathway scoring construction. Its node-score and pair-score pathways share a temporal encoder, and their scores are added at prediction time; the pair score combines a representation term with an explicit history residual. Pathway-Aligned Supervision (PAS) assigns random negatives to the node-score loss and historical negatives to the full pair-score loss, with random fallback when no eligible past partner exists. Crossing the two samplers with the two losses gives four assignments (Figure 1); comparing them tests directly whether choosing one sampler for the whole model misses a useful assignment.

(a) Four sampler assignments



R: random; H: historical/fallback.
The scoring architecture is fixed.

(b) TGAT: RH minus RR

Test MRR difference; seeds 42–44; mean ± sample SD

Dataset	Candidates	MRR change
tgbl-review	official	+0.2808 ±0.0020
tgbl-wiki	official	−0.0655 ±0.0256
tgbl-comment	official	+0.2598 ± 0.0237
tgbl-comment	uniform replay	−0.0472 ± 0.0020

The two comment rows reuse the same trained models.

Figure 1: Sampler assignment and ranking. (a) R/H selects each pathway's training negatives; RH is PAS. (b) Strict-TGAT paired $RH - RR$ test mean reciprocal rank (MRR), seeds 42–44 (mean ± sample SD). The two comment rows re-score the same models. Full candidate-specific comparisons are in Tables 2 and 4.

On tgbl-review, PAS reaches 0.4855 test mean reciprocal rank (MRR), compared with 0.2046 under all-random training. It also exceeds all-historical training and a validation-selected shared mixture, with the gains reproduced by disjoint confirmation seeds. Choosing only between shared samplers would therefore miss a useful assignment within the same predictor.

The same assignment lowers MRR on tgbl-wiki, and its advantage on tgbl-comment reverses when the same checkpoints are evaluated against uniform rather than official candidates. A history contrast can thus be available without improving the required ranking. Retraining and fixed-checkpoint score removal address a separate question: whether the fitted history component helps, which need not agree with the effect of changing its supervision.

Our contributions.

1. **A component-level learning condition.** We identify when a negative comparison gives an independently parameterized, deterministic history residual no direct gradient, while the shared representation can still learn.
2. **Pathway-aligned supervision.** PAS assigns negatives to scoring losses; a fixed-architecture factorial and shared-mixture control test what this assignment changes.
3. **The distinction between contrast and ranking benefit.** Independent confirmation, candidate replay and component experiments show when assignment gains persist, reverse, or differ from a fitted component's usefulness.

2 Assigning Negative Evidence to Scoring Pathways

Let $G = (\mathcal{V}, \mathcal{E}, \mathcal{T})$ be a temporal graph with interactions $(s, d, t) \in \mathcal{E}$. At query time t, we rank destinations using strict history $\mathcal{H}_t = \{(s, d, t_i) \in \mathcal{E} \mid t_i < t\}$. We hold the scoring construction fixed and change which negative scores enter each branch loss.

2.1 One Predictor, Two Supervised Scores

A temporal encoder f_θ produces node embeddings $\mathbf{z}_i = f_\theta(i, t)$. The additive scorer (Figure 2) is

$$\text{score}(s, d, t) = h_{\text{node}}(\mathbf{z}_s, \mathbf{z}_d, \mathbf{x}_e) + h_{\text{repr}}(P\mathbf{z}_s, P\mathbf{z}_d) + r_{\text{count}}(\mathbf{x}_p(s, d, t)), \tag{1}$$

Here, P is the pair projector; h_{node} includes a separate node projector with independent parameters. We write $s_N = g_{\text{node}} = h_{\text{node}}$ and $s_P = g_{\text{pair}} = h_{\text{repr}} + r_{\text{count}}$ for the node score and the full pair score, respectively. The representation term uses projected node embeddings, whereas the history residual uses three features of the ordered source–destination pair: log pair count, a seen-pair indicator, and pair recency. Both scores use the same temporal encoder; "node" and "pair" name the terminal scoring pathways rather than disentangled representations.

The node head also receives four strict-history activity statistics $\mathbf{x}_e$. The history residual receives $\mathbf{x}_p(s,d,t) = [\log(1+c_{\text{pair}}(s,d)), \mathbf{1}\{c_{\text{pair}}(s,d)>0\}, x_{\text{rec}}(s,d,t)]$, with all features computed from interactions strictly before t. Pair recency is $\log(1+t-t^{\text{last}}_{(s,d)})$ for a previously observed pair and -1 for an unseen pair. Appendix B.1 defines the node statistics and this pair-recency feature in full. Despite the identifier r_{count}, the residual is not count-only.

Two unseen pairs therefore give the residual the same complete input, $(0, 0, -1)$, even though their node embeddings can differ. This distinction connects the scoring construction to the analysis in Section 3: equal inputs cancel the independent history residual in a pairwise score difference, while the representation term can still distinguish the candidates and pass gradients to the shared encoder.

2.2 Branch-wise Losses and Negative Pools

The two negative pools are R and H. R draws uniformly with replacement from the distinct training destinations, without excluding the current target. H draws uniformly from the destinations with which the source has interacted strictly before the query time, excluding the current target. When no eligible historical destination remains, H makes an independent uniform fallback draw from the training-destination pool, again excluding the current target. We write "historical/fallback" to cover both cases. Each pair of random and historical negatives is reused by both heads, rather than sampled independently for each head.

For the main factorial with the full additive pair head, let $s_N = g_{\text{node}}$ and $s_P = g_{\text{pair}}$. With one negative from pool $S \in \{R, H\}$, the Bayesian Personalized Ranking (BPR) loss for branch b is

$$L_b^S = -\log \sigma\big(s_b(s,d^+,t) - s_b(s,d_S^-,t)\big), \qquad b \in \{N, P\}. \tag{2}$$

An assignment XY uses $L_N^X + \lambda L_P^Y$, with $\lambda = 0.25$; for PAS this is

$$\mathcal{L}_{\text{PAS}} = L_N^R + \lambda\, L_P^H, \qquad \lambda = 0.25. \tag{3}$$

Every candidate uses the shared encoder and both heads (Figure 2). Both losses can update the encoder; the pair loss reaches it through h_{repr}, while the history residual does not enter the node loss.

At inference, we add $s_N + s_P$ for the same candidate with unit weights and no learned gate. The training coefficient λ is therefore not an inference multiplier.

2.3 Factorial Assignment and the Shared-Mixture Control

In XY, the first letter selects the node-path sampler and the second the pair-path sampler: RR and HH share a sampler, RH is PAS, and HR reverses its allocation. The residual condition motivates H on the pair path without prescribing the node sampler, which the factorial tests.

Within each dataset and encoder, $RH - RR$ changes only the pair-path sampler while keeping the node sampler random. By contrast, $RH - HR$ compares two allocations that differ on both paths. The interaction $(RH - RR) - (HH - HR)$, measured in MRR, asks whether the pair-sampler effect depends on the node-path assignment; a positive interaction alone does not mean that RH improves over RR.

The shared-mixture comparator applies the same loss mixture to both pathways:

$$\begin{aligned} L_\omega &= (1-\omega)L_{RR} + \omega L_{HH}, \\ L_{RR} &= L_N^R + \lambda L_P^R, \quad L_{HH} = L_N^H + \lambda L_P^H, \quad \lambda = 0.25. \end{aligned} \tag{4}$$

Validation selects $\omega \in \{0, 0.25, 0.5, 0.75, 1\}$ (Appendix B.6). This loss weight ω differs from the probe's historical-sample fraction α. Gated and stop-gradient constructions are outside this factorial.

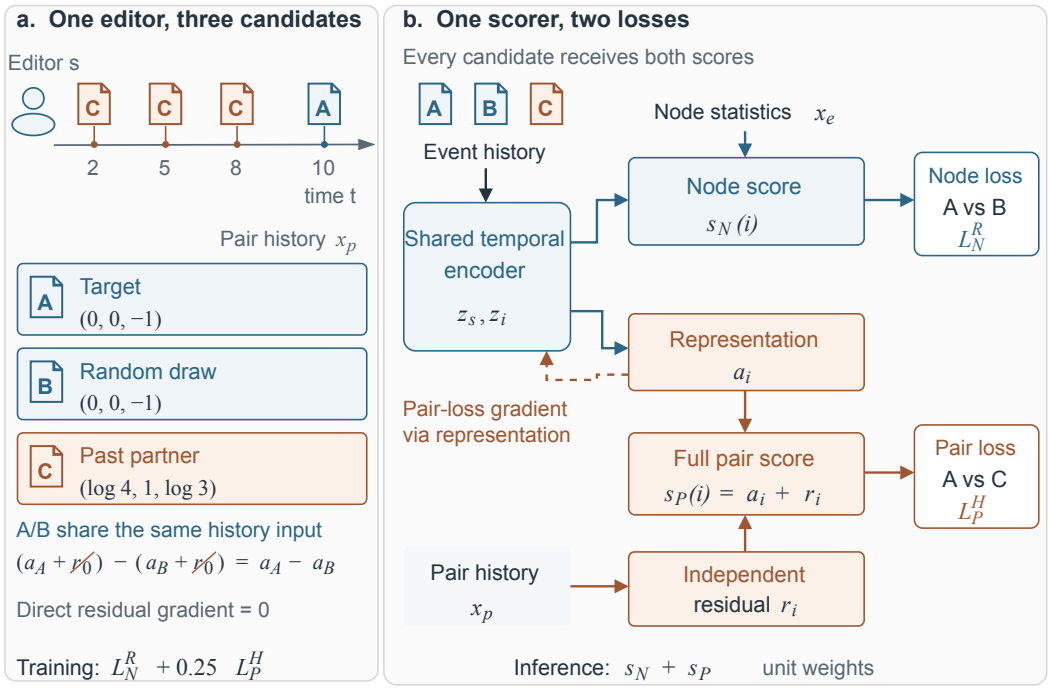


Figure 2: **Different comparisons within one shared scorer.** In this illustrative history, target A and one possible random negative B are unseen with editor s; C is a historical partner. Every candidate uses both scores. Let $a_i = h_{\text{repr}}(P\mathbf{z}_s, P\mathbf{z}_i)$ and $r_i = r_{\text{count}}(\mathbf{x}_p(s, i, t))$. PAS trains the node loss on A/B and the full pair loss on A/C. Identical A/B inputs cancel only the independent deterministic residual's direct parameter gradient under score-difference supervision; different A/C inputs do not guarantee a ranking benefit. The pair loss reaches the encoder through a_i. Training uses $L_N^R + 0.25L_P^H$; inference adds $s_N + s_P$ with unit weights.

3 When a Scoring Component Receives No Contrast

The explicit history residual illustrates how a sampler can leave one score component without training contrast. Write the score as $f(s, d) = g_{\text{rest}}(s, d, t; \theta_r) + r_{\text{count}}(\mathbf{x}_p(s, d, t); \theta_c)$, where the first term contains all other components.

Proposition 1 (Pair-history residual gradient cancellation)**.** *Suppose g_{rest} and the history inputs are independent of θ_c, and both candidates use the same deterministic residual module. For a differentiable pairwise-difference loss $\mathcal{L} = \ell(f(s, d^+) - f(s, d^-))$, including BPR (Rendle et al., 2009),*

$$\mathbf{x}_p(s, d^+, t) = \mathbf{x}_p(s, d^-, t) \quad \Longrightarrow \quad \nabla_{\theta_c}\mathcal{L} = \mathbf{0}. \tag{5}$$

In particular, two destinations unseen with the source have the same fixed history encoding and satisfy this condition.

Proof sketch. The equal residual outputs cancel for every θ_c, and the remaining score is independent of θ_c. Appendix A.1 gives the full derivation and the deterministic ReLU-subgradient case. □

The condition depends on the features seen by the added component, so it applies across scoring architectures with an independent, data-determined feature term. Fixed-temperature softmax losses share the cancellation when all candidate inputs coincide; when inputs are nearby, terminal-head gradients are also bounded under the stated regularity condition (Appendices A.4 and A.3). The representation term and shared encoder can still learn. This is a criterion for missing component-level contrast; the following experiments test its consequences for ranking.

3.1 Frequency of the Sufficient Event

In the matched strict-history measurement, both candidates are novel on 99.80% of tgbl-review training queries; a novel positive has an eligible historical alternative on 79.31%. The first event satisfies the residual-cancellation condition. The second measures available contrast, not recovered gradients. Both use all training queries as the denominator (Table 1; Appendix B.4).

Table 1: Strict-history sufficient-event audit; both percentages use all training queries. "Both novel" is sufficient for residual cancellation, not the complete zero-gradient event. Historical contrast means a novel positive has a strictly earlier alternative partner after current-target exclusion.

Dataset	Training queries	Both novel	Historical contrast
tgbl-review	3,413,837	99.80%	79.31%
tgbl-wiki	110,232	12.16%	6.37%

4 What Sampler Assignment Changes

4.1 Evaluation Setup

Datasets and encoders. Review and wiki carry the primary factorial and disjoint-seed confirmation; comment adds a complete factorial and same-model candidate replay. All are Temporal Graph Benchmark streams (Huang et al., 2023). TGAT (Xu et al., 2020) is primary; Section 5.1 substitutes TCL (Wang et al., 2021).

Training and selection. History inputs for a complete timestamp group are formed before the group is added to the state; optimization follows the original minibatch order. The factorial fixes architecture, three epochs and $\lambda = 0.25$; validation MRR selects checkpoints. Development seeds are 42–44; confirmation seeds 45–49 use the prespecified comparator. Appendix B.3 gives the model and sampling settings; Appendix B.9 gives the component and loss checks.

Evaluation candidates. Training pools R/H are distinct from evaluation candidates. We compare official candidates with fixed uniform replay (Section 4.4). Comment replay excludes the query target and all true targets at the same source/timestamp; Appendix B.8 gives its construction and composition.

Reporting. MRR averages reciprocal target ranks within the declared candidates. Strict cells and paired effects report mean $\pm$ sample SD; confirmation intervals quantify training-seed variation. Legacy tables retain their own conventions.

4.2 A Gain Missed by Shared Samplers

Table 2: Strict factorial test MRR under each dataset's official TGB candidates (models selected by validation MRR; seeds 42–44; mean $\pm$ sample standard deviation). In cell XY, X and Y select random (R) or historical/fallback (H) negatives for the node and pair paths.

	TGAT		TCL	
Cell	tgbl-review	tgbl-wiki	tgbl-review	tgbl-wiki
RR	0.2046 $\pm$0.0021	0.4198 $\pm$0.0424	0.2232 $\pm$0.0033	0.4624 $\pm$0.3486
RH (PAS)	0.4855 $\pm$0.0008	0.3544 $\pm$0.0645	0.4845 $\pm$0.0007	0.3940 $\pm$0.2796
HR	0.0685 $\pm$0.0055	0.4516 $\pm$0.0387	0.1085 $\pm$0.0010	0.4754 $\pm$0.3172
HH	0.1253 $\pm$0.0042	0.3693 $\pm$0.0735	0.2239 $\pm$0.0054	0.4010 $\pm$0.2860

With the node sampler fixed to R, TGAT's $RH - RR$ is $+0.2808 \pm 0.0020$ on review and -0.0655 ± 0.0256 on wiki (Table 2). On review, RH exceeds RR and HH exceeds HR: changing the pair sampler helps under either node sampler. Yet HH is below RR, so comparing shared samplers misses the gain.

4.3 Comparison with a Selected Shared Mixture

On review, RH exceeds the validation-selected shared mixture by $+0.0477 \pm 0.0252$ MRR. Validation selects $\omega = 0.25$, with test MRR 0.4378 ± 0.0250 for that comparator. On wiki, validation selects the all-random endpoint ($\omega = 0$), and RH remains adverse.

Disjoint confirmation seeds 45–49 reproduce the two directions: $RH - RR$ is $+0.2793 \pm 0.0045$ on review and -0.0493 ± 0.0261 on wiki. RH also exceeds the preselected review mixture by $+0.0541 \pm 0.0140$. The review pair-sampler contrast has a two-sided 95% training-seed confidence interval of $[+0.2737, +0.2849]$; every review confirmation seed meets the prespecified positive-interaction and positive-mixture-margin criteria. Appendices B.5–B.7 give the complete contrasts, grid, and confirmation results.

Table 3: Disjoint-seed confirmation (45–49): paired MRR effects, mean ± sample SD. The preselected wiki mixture is RR ($\omega = 0$). Full contrasts and confidence intervals are in Table 7.

Contrast	tgbl-review	tgbl-wiki
$RH - RR$	$+0.2793 \pm 0.0045$	-0.0493 ± 0.0261
$RH-$ preselected mixture	$+0.0541 \pm 0.0140$	-0.0493 ± 0.0261

4.4 Ranking Gains Depend on the Candidate Set

Under the official candidates, comment provides a second positive assignment result: PAS reaches 0.8894 MRR, with a paired $RH - RR$ gain of $+0.2598 \pm 0.0237$ (Table 4A). The other two simple effects, $HH - HR$ and $HR - RR$, do not have consistent signs across the three seeds, so the full factorial does not establish a uniform benefit for every sampler change.

We then re-score the same 12 validation-selected checkpoints against a fixed, model-blind uniform slate of 20 negatives per query. No model is retrained or reselected. $RH - RR$ becomes -0.0472 ± 0.0020, and $RH - HR$ also reverses sign. The change is therefore a reversal in the relative value of training assignments for the same fitted models, rather than a consequence of fitting new models to a different evaluation protocol.

Review shows a complementary result (Table 4B). Its same frozen RR/RH checkpoints are evaluated on 728,919 queries with 100 negatives per query in each slate. Under fixed uniform candidates, RR and RH reach 0.4471 ± 0.0019 and 0.4487 ± 0.0012 MRR, respectively. Their paired difference is $+0.0016 \pm 0.0008$, compared with $+0.2808 \pm 0.0020$ under the official candidates. Thus the large review advantage nearly disappears, whereas the comment advantage changes sign. Absolute MRRs across candidate protocols are not treated as model improvements; the comparison is the paired assignment effect within each protocol.

The candidate populations clarify how the ranking task changed. On comment, strictly historical destinations occupy 32.17% of official negative slots and 0.114% of uniform slots, measured over the same 132,943,520 slots. On review, the corresponding mean counts are 7.86 and 0.0043 historical negatives per 100-negative list. These joint changes show that the benefit of assignment depends on the evaluation alternatives, without isolating historical-candidate frequency as the sole cause. A positive factorial interaction can still coexist with a negative $RH - RR$ effect, as in comment's uniform replay; it should not be read as a PAS gain over all-random training. Appendix B.8 reports the candidate construction and complete comment contrasts.

4.5 What the History Residual Contributes

We distinguish retraining the residual under different supervision from removing its score after training.

Component retraining. Component-restricted retraining compares a pair path retaining only the explicit pair-history residual with one retaining only the representation term. On review, $RH - RR$ is $+0.2790 \pm 0.0011$ and $+0.0034 \pm 0.0022$, respectively, so the assignment effect is much larger in the history-only variant. In that variant, the pair loss trains the residual but has no path to the temporal

Table 4: Candidate replay with fixed trained models. A: tgbl-comment uses 20 negatives per query and all four training assignments. B: tgbl-review uses 100 negatives per query and the recorded RR/RH checkpoints. Means and sample standard deviations use seeds 42–44; assignment differences are paired within seed. Models are neither retrained nor reselected. The two datasets retain separate query populations and candidate protocols; comparisons concern assignment effects within each dataset and protocol.

A. tgbl-comment: four training assignments

Quantity	Official candidates	Fixed uniform
RR	0.6296 ± 0.0226	0.9156 ± 0.0017
RH (PAS)	0.8894 ± 0.0018	0.8684 ± 0.0024
HR	0.6275 ± 0.0242	0.9084 ± 0.0011
HH	0.6170 ± 0.0128	0.8319 ± 0.0020
$RH - RR$	$+0.2598 \pm 0.0237$	-0.0472 ± 0.0020

B. tgbl-review: RR/RH; 728,919 queries; 100 negatives per query

Candidates	RR MRR	RH MRR	Paired $RH - RR$	Mean historical negatives per list
Official	0.2046 ± 0.0021	0.4855 ± 0.0008	0.2808 ± 0.0020	7.86
Fixed uniform	0.4471 ± 0.0019	0.4487 ± 0.0012	0.0016 ± 0.0008	0.0043

encoder. These retrained variants test component dependence of the assignment effect, without identifying a training mediator.

Table 5: Component retraining (A) versus fixed-checkpoint residual removal (B), seeds 42–44, mean $\pm$ sample SD. B fixes models, queries, candidates and history and measures $\Delta_C = \mathrm{MRR}(E + R + C) - \mathrm{MRR}(E + R)$, with $E = h_{\mathrm{node}}$, $R = h_{\mathrm{repr}}$, $C = r_{\mathrm{count}}(\mathbf{x}_p)$.

A. Component retraining (tgbl-review)

Pair pathway	$RH - RR$	Pair loss $\rightarrow$ encoder
Representation + history residual	$+0.2808 \pm 0.0020$	Yes
History residual only	$+0.2790 \pm 0.0011$	No
Representation only	$+0.0034 \pm 0.0022$	Yes
History effect $-$ representation effect: $+0.2756 \pm 0.0027$		

B. History-residual removal at fixed checkpoints (no retraining)

Dataset / candidates	RR Δ_C	RH Δ_C
review / official	-0.1272 ± 0.0023	$+0.1509 \pm 0.0017$
review / fixed uniform	$+0.000050 \pm 0.000013$	-0.000974 ± 0.000004
wiki / official	$+0.2627 \pm 0.0439$	$+0.1974 \pm 0.0661$

Table 5B uses review's replay populations (Table 4B) and 23,621 wiki queries with 998–999 negatives. Review's residual contribution changes sign between RR and RH on official candidates; both uniform-candidate increments have magnitude below 0.001. On wiki it is positive within both arms although RH training lowers full-model MRR. Component usefulness need not imply a benefit from changing its supervision (full decomposition in Appendix B.10).

The size of the review effect becomes clearer after removing the fitted residual. Subtracting the reported mean increments from the full-model means leaves approximately 0.332 MRR for RR and 0.335 for RH under official candidates. The much larger full-model gap accompanies a residual contribution that changes from harmful under RR to useful under RH. On wiki, the corresponding residual-free means are both approximately 0.157, while the residual contributes more under RR. These are arithmetic decompositions of fixed predictions, not estimates of a causal training mediator. Together with component retraining, they locate the contrast more precisely than the full-model scores alone.

5 BEYOND THE PRIMARY CONSTRUCTION

5.1 ENCODER AND LOSS CHECKS

Encoder substitution. Replacing TGAT with Transformer-based Dynamic Graph Modelling via Contrastive Learning (TCL), while keeping the constructed head and BPR loss fixed, gives the same mean directions on review and wiki (Table 2). Wiki has large seed variation; Appendix B.9 reports the available paired contrasts. TCL's native contrastive objective is outside this experiment.

Multi-negative loss. A separate InfoNCE cohort uses $K \in \{1, 5\}$ negatives, temperature 1.0 and seeds 50–52. Review's $K = 5$ $RH - RR$, $RH - HR$ and interaction are positive in every seed; wiki's mean $RH - RR$ is adverse at both K. The paired change in review's $RH - HR$ between $K = 1$ and 5 is small, providing no evidence that increasing K improves assignment effects.

For review at $K = 5$, $RH - RR$ is $+0.2814 \pm 0.0016$; the corresponding wiki effect is -0.0252 ± 0.0222. Full numerical contrasts and cohort-specific protocols are in Appendix B.9 (Table 9).

5.2 WHAT NODE AND PAIR INFORMATION PREDICT

The diagnostic probe asks when node and pair information are useful to a predictor, complementing the strict study of which loss receives which negatives. Under its record-order history rule, which permits same-timestamp reads, full DTLP scores below entity-only on review (0.223 versus 0.361 MRR), while pair-only nearly matches full DTLP on wiki (0.698 versus 0.721). Mixing historical negatives raises review to 0.497 but lowers wiki to 0.624. These results show that making pair history available and learning a useful combined score are separate issues. They motivate the strict factorial but do not replicate its architecture or timestamp-group protocol. Appendices C–D retain the feature ablations, sampler sweeps, and score-distribution diagnostics.

The TPNet-based experiments extend node/pair prediction to five streams (Table 36); their per-stream constructions and reference models differ from the strict predictor. The node-property study separately tests history statistics and graph features: DTLP-node is competitive on reddit and genre but trails strong heuristics on trade and token (Table 48). These studies show complementary uses of node and pair histories beyond the primary link-ranking experiment.

5.3 ADDING HISTORY TO AN EXISTING SCORE

The CRAFT studies separate choosing negatives for an existing score from adding a new history pathway. Model-wide sampler mixing in CRAFT's single score helps wiki but hurts review and comment. In a separate retrofit, adding a residual and gate to a fixed backbone improves both reported datasets; a label-decorrelated placebo returns to the wiki backbone. A single-seed uniform-negative residual also improves wiki (0.7697 versus 0.7042), so these results support an added pathway without establishing that historical negatives are necessary. No review placebo is available. Appendix E gives the constructions and controls, which differ from jointly trained PAS.

6 DISCUSSION AND RELATED WORK

A useful fitted component and a useful training assignment are different results. PAS improves ranking on review, whereas on wiki the history residual helps both fitted models even though PAS lowers full-model MRR. Candidate replay adds a second distinction: an assignment's advantage can change when the same model faces different alternatives. Providing training contrast, learning a useful score component and improving the required ranking must therefore be assessed separately.

The transferable design question is which inputs each trainable score actually contrasts. For an independently parameterized additive component with data-defined inputs, a negative may be distinguishable by the rest of the predictor while presenting the same input to that component. The local cancellation result makes this missing contrast explicit. It motivates choosing sampler–component

pairs, rather than labeling a negative as hard or easy for the model as a whole. The factorial, shared-mixture control and component studies then test whether the resulting assignment helps the intended ranking problem.

PAS requires separately addressable terminal scores and losses, not separate encoders. A shared encoder may receive both losses through the representation pathways. Model-wide mixing in a single-path scorer and an external residual fitted to a fixed backbone are different design choices; Appendix E reports them separately. The broader degree-negative control also distinguishes harder comparisons from useful pair contrast: it increases training loss without reproducing the historical-negative pair/entity gradient response (Table 13). Historical partners can interact with the source again later while remaining valid negatives for the instantaneous prediction task. The overlap diagnostics in Appendix C.7 describe this structure of the candidate pools; they do not isolate why PAS harms wiki. Recurrence and later reappearance therefore motivate checking an assignment under the intended candidates, rather than selecting a sampler by a universal threshold.

The strict comparison fixes the pair-loss weight and three-epoch budget. Assigning samplers changes no parameter count, but maintaining exact pair history requires state that grows with the number of distinct directed pairs. The reported timings are co-scheduled measurements, and strict-PAS peak VRAM was not recorded (Appendix H). Native memory-based adaptations and bounded-history approximations remain outside the evaluated constructions.

Temporal graph learning and benchmarks. Memory, temporal-neighborhood and sequence models represent interaction history in different ways (Rossi et al., 2020; Xu et al., 2020; Yu et al., 2023; Lu et al., 2024). PAS concerns the negative evidence entering terminal scores, independently of that encoder choice. Benchmark studies show sensitivity to candidates (Huang et al., 2023; Gastinger et al., 2024); TGB-Seq emphasizes dynamics beyond repeated edges (Yi et al., 2025b). Our fixed-model replay tests whether candidate changes reverse a training intervention's value.

Model-level sampling and pathway assignment. Graph recommendation and knowledge-graph embedding commonly use hard negatives (Ying et al., 2018; Sun et al., 2019; Ahrabian et al., 2020; Huang et al., 2021; Zhang et al., 2019). Debiased contrastive and positive-unlabeled methods reweight losses or candidates to correct likely false negatives (Chuang et al., 2020; Robinson et al., 2021; Elkan & Noto, 2008). DICE trains additive interest/conformity components on cause-specific, popularity-conditioned subsets (Zheng et al., 2021); CurNM uses temporal factor disentanglement to select from random, historical and hard pools (Chen et al., 2024). PAS instead tests random versus historical/fallback assignments within one separable temporal predictor.

Local cancellation and feature competition. Gupta & Bedathur (2024) study model-level sampling difficulty; we isolate exact pair-history residual parameter-gradient cancellation (Proposition 1). Pezeshki et al. (2021) study gradient starvation under feature competition. Recurrence baselines (Poursafaei et al., 2022; Liu et al., 2022; Li et al., 2021; 2022; Trivedi et al., 2017) show that historical signals can be useful. We test when changing a pair score's supervision improves ranking.

7 Conclusion

Negative supervision can be designed for individual scoring pathways. PAS yields a review gain beyond the tested shared samplers, while the harm on wiki and the candidate-dependent reversal on comment distinguish useful training contrast from ranking benefit. The local analysis identifies when a component receives no contrast; assignment and component experiments test whether changing it helps. Pathway assignments should therefore be compared under the intended evaluation candidates.

AI USE STATEMENT

The research questions, methods, experimental design and scientific conclusions were developed by the authors. Generative AI tools assisted the organization and presentation of research materials, language editing and translation, figure preparation, and checks of experimental and analytical rigor, including related supporting code. These tools were used under author direction; scientific judgments and responsibility for the final manuscript remain with the authors.

REPRODUCIBILITY STATEMENT

The proposition states its conditions in Section 3, with the full proof in Appendix A.1. For the strict main evidence, start with Section 2 and Appendix B: timestamp semantics and event counts, the complete factorial and shared mixture, disjoint confirmation, same-model comment replay, and encoder/component/loss checks. Tables 2 and 4 give the main results; Appendix B.10 separately reports fixed-model score contributions. The anonymous supplement's current table map links these objects to implementations, frozen protocols, sealed summaries and existing verification materials.

The supplement retains the separable objective, strict timestamp handling, exact source/input identifiers, aggregates recorded before manuscript analysis, independent recomputation scripts and representative tests. It does not include all datasets, checkpoints, training logs or per-query matrices. Probe defaults are in Appendix C.2, the distinct PDRE algorithm in Appendix E.2, and other candidate protocols beside their benchmark tables. Anonymous materials sufficient to reproduce the held-fixed CRAFT retrofit are not included, so those contextual cells cannot be reproduced from the submitted supplement alone. The earlier dual-path comparison in Table 29 uses a different construction and omits HR. The strict four-assignment study and disjoint-seed confirmation are subsequent comparisons; internal artifact-group labels do not identify the experiments in that earlier table.

ETHICS STATEMENT

This study uses public temporal-graph benchmarks (TGB, TGB 2.0, DyGLib, and TGB-Seq) and collects no human-subject data. The submission materials are anonymous.

REFERENCES

Kian Ahrabian, Aarash Feizi, Yasmin Salehi, William L. Hamilton, and Avishek Joey Bose. Structure aware negative sampling in knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6093–6101, Online, 2020. Association for Computational Linguistics.

Ziyue Chen, Tongya Zheng, and Mingli Song. Curriculum negative mining for temporal networks. *arXiv preprint arXiv:2407.17070*, 2024.

Ching-Yao Chuang, Joshua W. Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *International Conference on Learning Representations*, 2023.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220. ACM, 2008.

Or Feldman, Krishna Sri Ipsit Mantri, Moshe Eliasof, and Chaim Baskin. Revisiting node affinity prediction in temporal graphs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Julia Gastinger, Shenyang Huang, Mikhail Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt, Reihaneh Rabbany,

and Guillaume Rabusseau. TGB 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 140199–140229, 2024.

Shubham Gupta and Srikanta Bedathur. Robust training of temporal GNNs using nearest neighbour based hard negatives. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data*, pp. 54–63. ACM, 2024.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2056–2073, 2023.

Tinglin Huang, Yuxiao Dong, Ming Ding, Zhen Yang, Wenzheng Feng, Xinyu Wang, and Jie Tang. MixGCF: An improved training method for graph neural network-based recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 665–674. ACM, 2021.

Ce Li, Rongpei Hong, Xovee Xu, Goce Trajcevski, and Fan Zhou. Simplifying temporal heterogeneous network for continuous-time link prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 1288–1297. ACM, 2023.

Zixuan Li, Xiaolong Jin, Wei Li, Saiping Guan, Jiafeng Guo, Huawei Shen, Yuanzhuo Wang, and Xueqi Cheng. Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 408–417. ACM, 2021.

Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. Complex evolutional pattern learning for temporal knowledge graph reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 290–296. Association for Computational Linguistics, 2022.

Yushan Liu, Yunpu Ma, Marcel Hildebrandt, Mitchell Joblin, and Volker Tresp. TLogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(4):4120–4127, 2022.

Xiaodong Lu, Leilei Sun, Tongyu Zhu, and Weifeng Lv. Improving temporal link prediction via temporal walk matrix projection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 141153–141182, 2024.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32928–32941, 2022.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461. AUAI Press, 2009.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on Graph Representation Learning*, 2020.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019.

Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3462–3471. PMLR, 2017.

Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. TCL: transformer-based dynamic graph modelling via contrastive learning. *CoRR*, abs/2105.07944, 2021. URL https://arxiv.org/abs/2105.07944.

Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations*, 2020.

Lu Yi, Runlin Lei, Fengran Mo, Yanping Zheng, Zhewei Wei, and Yuhang Ye. Future link prediction without memory or aggregation. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Lu Yi, Jie Peng, Yanping Zheng, Fengran Mo, Zhewei Wei, Yuhang Ye, Zixuan Yue, and Zengfeng Huang. TGB-Seq benchmark: Challenging temporal GNNs with complex sequential dynamics. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 974–983. ACM, 2018.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67686–67700, 2023.

Yongqi Zhang, Quanming Yao, Yingxia Shao, and Lei Chen. NSCaching: Simple and efficient negative sampling for knowledge graph embedding. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 614–625. IEEE, 2019.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Depeng Jin, and Yong Li. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*. ACM, 2021.

SUPPLEMENTARY MATERIAL

Appendix reading map. Appendix A gives the mathematical conditions; B, strict protocols and extended results; C, the diagnostic probe followed by gradient and historical-negative diagnostics; D, probe ablations; E, PDRE and external retrofits; F, results by task and protocol; G, retrospective recurrence context; and H, resources and tuning. Each retains its construction and comparison target.

A EXACT GRADIENT RESULTS AND THEIR SCOPE

This appendix collects the proof and terminal-head extensions of Section 3. The legacy feature-scope distinction is documented with the probe in Appendix C.3.

A.1 PROOF OF PROPOSITION 1

Let $f(s,d,t) = g_{\text{rest}}(s,d,t;\theta_r) + r_{\text{count}}(\mathbf{x}_p(s,d,t);\theta_c)$ and $\mathcal{L} = \ell(f(s,d^+,t) - f(s,d^-,t))$ for differentiable ℓ. The feature map $\mathbf{x}_p$ is independent of θ_c, and both candidates use the same deterministic residual module. Define $\Delta f = f(s,d^+,t) - f(s,d^-,t)$. Then

$$\nabla_{\theta_c}\mathcal{L} = \ell'(\Delta f)\nabla_{\theta_c}\left[r_{\text{count}}(\mathbf{x}_p(s,d^+,t);\theta_c) - r_{\text{count}}(\mathbf{x}_p(s,d^-,t);\theta_c)\right]. \tag{6}$$

When the positive and negative residual inputs are identical, the bracketed difference is identically zero for every θ_c, so its derivative is exactly zero. If neither candidate has prior interaction with source s, both vectors equal the fixed novel-pair encoding $\mathbf{x}_\emptyset$, which is one such identical-input instance. Independently sampled dropout, candidate-specific residual modules, or features that depend on θ_c are outside the proposition.

The proposition is stated at differentiable points. For a ReLU residual, identical inputs follow the same activation pattern, and automatic differentiation applies the same deterministic subgradient to both evaluations. Their residual-parameter gradients therefore cancel as well.

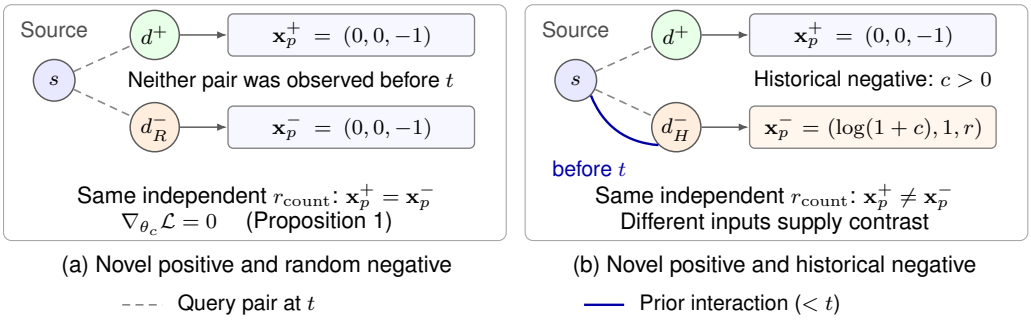


Figure 3: Two local history-input contrasts. (a) A novel positive and a novel random negative both have residual input $(0,0,-1)$, giving residual-parameter cancellation under Proposition 1. (b) An eligible historical negative supplies a different residual input. The representation pathway and encoder can still receive gradients; input contrast alone does not guarantee improved ranking.

A.2 FREQUENCY OF THE SUFFICIENT EVENT

For query $q = (s,d^+,t)$, let $\pi_q(d)$ denote the implemented negative-sampling distribution over candidate set $\mathcal{C}_q$. Let $Q(\cdot \mid s)$ denote the query distribution conditional on source s. The source-conditional probability of the sufficient both-novel event is

$$P_s^{\text{both-novel}} = \mathbb{E}_{q\sim Q(\cdot|s)}\left[\mathbf{1}\{d^+ \notin \mathcal{H}_s(t)\}\sum_{d\in\mathcal{C}_q\setminus\mathcal{H}_s(t)}\pi_q(d)\right]. \tag{7}$$

Under uniform sampling from $\mathcal{C}_q$, the inner sum is $1 - |\mathcal{H}_s(t)\cap\mathcal{C}_q|/|\mathcal{C}_q|$. If A_q is the positive-novelty indicator and m_q the sampler probability mass, then $\mathbb{E}[A_q m_q] = \mathbb{E}[A_q]\mathbb{E}[m_q] + \text{Cov}(A_q, m_q)$. Thus factorization requires zero covariance; it is not an identity.

A.3 A Conditional Bound for Terminal Pair Heads

Exact residual-parameter cancellation holds under identical residual inputs. Identical inputs are sufficient, not necessary, for a zero residual gradient. A nearby-input statement is possible for terminal head parameters under explicit regularity assumptions, but it must not be propagated to the projector or backbone.

Proposition 2 (Terminal-head gradient bound)**.** *Let $f^{\pm} = a^{\pm} + h(u^{\pm}; \phi)$, where $a^{\pm}$ and $u^{\pm}$ are independent of ϕ, and let h be differentiable. Assume $u \mapsto \nabla_{\phi} h(u; \phi)$ is L_g-Lipschitz. For BPR,*

$$\|\nabla_{\phi} \mathcal{L}_{\mathrm{BPR}}\| \leq \sigma(-\Delta f) L_g \left\|u^{+} - u^{-}\right\|. \tag{8}$$

Proof. The chain rule gives, up to sign,

$$\|\nabla_{\phi} \mathcal{L}_{\mathrm{BPR}}\| = \sigma(-\Delta f) \left\|\nabla_{\phi} h(u^{+}; \phi) - \nabla_{\phi} h(u^{-}; \phi)\right\|. \tag{9}$$

Applying the assumed Lipschitz condition proves Eq. 8. □

This is only an upper bound for terminal parameters ϕ. A ReLU head need not satisfy the gradient-Lipschitz assumption across activation boundaries, so the statement applies only in regions where the assumption holds. It does not apply when u depends on ϕ, does not cover projector or backbone gradients, and does not prove that historical negatives recover a useful direction.

A.4 InfoNCE boundary

Proposition 1 is stated for pairwise-difference losses on a separately parameterized count residual. Separately, an exact terminal fixed-input identity holds for fixed-temperature softmax cross-entropy. Let $z_i = (a_i + h_{\phi}(x_i))/\tau$ be the positive and K negative logits, where a_i and residual inputs x_i are independent of terminal parameters ϕ and $\tau > 0$ is fixed. Then

$$\nabla_{\phi} \mathcal{L}_{\mathrm{NCE}} = \frac{1}{\tau} \sum_{i=0}^{K} (p_i - \mathbb{1}[i = 0]) \nabla_{\phi} h_{\phi}(x_i). \tag{10}$$

If every x_i is identical, the terminal gradients are identical and the coefficients sum to $\sum_i p_i - 1 = 0$. This identity holds for any K and fixed temperature, but it does not cover the projector, backbone, or a learned temperature.

No universal collision threshold follows. After projecting updates onto a fixed scalar direction, suppose a helpful event contributes magnitude $g_+ > 0$ and an interfering event contributes $g_- > 0$. If interference occurs with probability p, the expectation $(1 - p)g_+ - pg_-$ crosses zero at $p^{\star} = g_+/(g_+ + g_-)$. The value is $1/2$ only in the symmetric case $g_+ = g_-$. With non-collinear updates, even this scalar threshold need not exist.

B Strict PAS Construction and Supporting Checks

This appendix gives the protocols and extended results for the strict sampler-assignment study. It separates the matched-history event audit, factorial and shared-mixture comparisons, disjoint confirmation, same-checkpoint candidate replay, and encoder/component/loss checks. The fixed-checkpoint score decomposition is reported as a diagnostic rather than a retraining effect. Each unit retains its own cohort and comparison target; results from the legacy probes and alternative pathway constructions are documented separately.

B.1 Strict-History Scoring Features

The node head also uses four strict-history statistics, $\mathbf{x}_e = [\log(1 + n_s), \log(1 + n_d), r_s, r_d]$. Here n_v counts prior appearances of node v as either endpoint, including repeated interactions, and $r_v = \log(1 + t - t_v^{\mathrm{last}})$ measures elapsed time since its latest prior interaction with any partner, with $r_v = -1$ for an unseen node.

The pair-history residual's pair-recency feature (PRF) is

$$x_{\mathrm{rec}}(s,d,t) = \begin{cases} \log\left(1 + t - t^{\mathrm{last}}_{(s,d)}\right), & \text{if } (s,d) \text{ was seen before } t, \\ -1, & \text{otherwise,} \end{cases} \tag{11}$$

where $t^{\mathrm{last}}_{(s,d)}$ is the most recent prior interaction of that ordered pair.

B.2 Strict History and Timestamp Ties

For each timestamp group, the history statistics for positive and sampled candidates are materialized using only interactions strictly before t. The group is committed to the history state only after these inputs have been formed. Training then processes the materialized records in the original minibatch order (Appendix B.3). The history barrier enforces $t^{\max}_{\mathrm{used}} < t$; it does not impose one optimizer step per timestamp or imply invariance of the entire optimization trajectory to within-group reordering.

This rule differs from the earlier record-order audit, which checked only $t^{\max}_{\mathrm{used}} \leq t$ and found same-timestamp reads on 100% / 34.1% / 3.8% of sampled review/comment/wiki queries. That condition permits dependence on within-timestamp order and does not establish $t_i < t$. The strict rule supersedes it for the primary assignment evidence. The older record-order DTLP, PDRE, count-filter and broad benchmark results use their own protocols and do not replicate the strict experiment.

B.3 Model and Optimization Settings

The primary review/wiki TGAT cohort uses one temporal-attention layer, two heads and the ten most recent neighbors. The TCL replacement uses two layers, two heads and twenty recent neighbors. Both output 172-dimensional embeddings, use 100-dimensional time encodings and encoder dropout 0.1. Their BPR heads have the same structure:

Component	Dimensions and activation
Node projector	Linear, $172 \to 86$
Pair projector (separate parameters)	Linear, $172 \to 86$
Node score	$176 \to 172 \to 1$, ReLU hidden layer
Pair representation score	$172 \to 172 \to 1$, ReLU hidden layer
History residual	$3 \to 172 \to 1$, ReLU hidden layer

Each projector is shared between the source and destination within its own pathway. The node-score input concatenates two projected embeddings and four activity statistics; the pair-representation input concatenates its two projected embeddings. The three heads have linear outputs and no head dropout. The encoder and heads are optimized jointly by Adam at learning rate 10^{-4}, zero weight decay, non-finite gradients replaced by zero before gradient-norm clipping at 5, and batches of 400 positives in stream order. The budget is three epochs and $\lambda = 0.25$. Validation MRR is evaluated after each epoch; a later epoch replaces an exactly tied best checkpoint. The review evaluation batches contain 400 queries and wiki batches 200. The TGAT and TCL negative-forward microbatch limits are 16,000 and 1,024, respectively.

Sampling and state updates. Random destinations are drawn with replacement from unique training destinations, without excluding the current target or applying a query-time birth filter. Historical sampling uses the source's strictly earlier partners and excludes the current target. For K negatives, it samples without replacement when the historical pool has at least K members and with replacement otherwise. An empty historical pool falls back to random destinations excluding the target. Neither training rule excludes all other targets at the same timestamp; comment's evaluation-slate exclusion is separate.

Each query has one generated random set and one historical/fallback set, reused by the two heads; the four assignments choose which set enters each loss. At every training epoch, history statistics and the historical sampler's random state are reset, while model and optimizer states persist. Complete timestamp groups are processed to materialize strictly earlier-history features before optimization in the original minibatch order. This history barrier does not impose one optimizer step per timestamp. The random sampler continues its run-level random stream across epochs. For testing, the selected

model rebuilds training and validation history before advancing through the test stream. The additional cohorts retain the seed, loss and continuation settings in their respective protocol sections.

B.4 Strict Sufficient-Event Audit

Proposition 1 makes identical residual inputs a sufficient condition, so we audit the both-novel subevent directly, under the same strict timestamp-group implementation and random-destination pool as the controlled factorial. Table 1 reports exhaustive counts over official training queries, not a proxy correlation across incompatible protocols.

B.5 Extended Factorial Contrasts

Table 6 retains the full TGAT contrasts reported in the main study (official candidates, seeds 42–44). Review interaction is positive in every seed; validation selects $\omega = 0.25$, whose test MRR is given in the mixture grid below. Wiki selects RR ($\omega = 0$) by validation; both principal contrasts are negative.

Table 6: TGAT paired contrasts; seeds 42–44, mean ± sample SD.

Dataset / construction	Quantity	MRR statistic
review	$RH - RR$	$+0.2808 \pm 0.0020$
review	$RH - HR$	$+0.4170 \pm 0.0062$
review	$(RH - RR) - (HH - HR)$	$+0.2240 \pm 0.0017$
review	RH minus selected mixture	$+0.0477 \pm 0.0252$
wiki	$RH - RR$	-0.0655 ± 0.0256
wiki	$RH - HR$	-0.0973 ± 0.0463

B.6 Global-Mixture Comparator Grid

The loss mixture in Eq. (4) is evaluated at interior weights $\omega \in \{0.25, 0.5, 0.75\}$; Table 2 gives its RR and HH endpoint conditions. The strict factorial's fixed three-epoch budget is a compute boundary; longer budgets were not evaluated.

Across the tuned global-mixture grid, test MRR is 0.4378 ± 0.0250 at $\omega = 0.25$, 0.3195 ± 0.0285 at $\omega = 0.5$, and 0.2138 ± 0.0041 at $\omega = 0.75$ on review, and 0.4124 ± 0.0589, 0.4098 ± 0.0463, and 0.3827 ± 0.0600 at the same three weights on wiki. Validation selects review: $\omega = 0.25$; wiki: RR ($\omega = 0$).

B.7 Disjoint-Seed Confirmation

Table 7: Shared-mixture comparison and disjoint-seed confirmation. Validation MRR selects the shared loss weight in development; confirmation reuses that comparator. Entries are means ± sample SD, with 95% training-seed confidence intervals shown separately.

Quantity	tgbl-review	tgbl-wiki
Development seeds 42–44		
Selected ω	0.25	0 (RR)
Selected shared-mixture MRR	0.4378 ± 0.0250	0.4198 ± 0.0424
RH − selected mixture	$+0.0477 \pm 0.0252$	-0.0655 ± 0.0256
Confirmation seeds 45–49		
$RH - RR$	$+0.2793 \pm 0.0045$	-0.0493 ± 0.0261
95% CI	$[+0.2737, +0.2849]$	$[-0.0817, -0.0168]$
$RH - HR$	$+0.4145 \pm 0.0034$	-0.0586 ± 0.0369
$(RH - RR) - (HH - HR)$	$+0.2211 \pm 0.0051$	$+0.0097 \pm 0.0188$
95% CI	$[+0.2147, +0.2275]$	$[-0.0137, +0.0331]$
RH − preselected mixture	$+0.0541 \pm 0.0140$	-0.0493 ± 0.0261
95% CI	$[+0.0367, +0.0715]$	$[-0.0817, -0.0168]$

Seeds 45–49 were launched only after the comparator was specified using development seeds 42–44. Table 7 retains the confirmation contrasts; the CI is a two-sided 95% Student-t interval over training-seed variation. Every review seed clears both prespecified criteria: positive interaction and a positive margin over the preselected shared mixture. Wiki remains adverse, strengthening dataset conditionality rather than a protocol-robust method claim.

Table 7 reports the complete confirmation contrasts and confidence intervals.

B.8 Comment Replay Candidate Composition

Table 8: Prespecified tgbl-comment condition means and paired effects under the benchmark's own negative-candidate lists versus uniform replay of the same trained models (seeds 42–44; mean $\pm$ sample standard deviation). Candidate sets differ across the two columns, so absolute MRRs are not compared across protocols; only paired arm effects within each column are interpreted.

Quantity	**Benchmark candidate lists**	**Uniform replay**
RR	0.6296 ± 0.0226	0.9156 ± 0.0017
RH (PAS)	0.8894 ± 0.0018	0.8684 ± 0.0024
HR	0.6275 ± 0.0242	0.9084 ± 0.0011
HH	0.6170 ± 0.0128	0.8319 ± 0.0020
$RH - RR$	$+0.2598 \pm 0.0237$	-0.0472 ± 0.0020
$HH - HR$	-0.0105 ± 0.0314	-0.0766 ± 0.0012
$HR - RR$	-0.0021 ± 0.0053	-0.0072 ± 0.0016
$(RH - RR) - (HH - HR)$	$+0.2703 \pm 0.0164$	$+0.0293 \pm 0.0030$
$RH - HR$	$+0.2619 \pm 0.0256$	-0.0400 ± 0.0032

For the separate tgbl-comment uniform replay of Section 4.4, the slate draws 20 negatives per query uniformly, with a fixed seed and without using model scores, from the unique training-split destinations, excluding only the query's own target and same-timestamp positives. In the supplied paired composition audit, 32.17% of official negative candidates and 0.114% of fixed-uniform candidates are strictly historical. Both rates use 132,943,520 candidate slots from the same 6,647,176 test queries with 20 negatives each. This describes the changed candidate population; it does not identify which part of that change causes the routing reversal.

B.9 Encoder, Component, and Loss Checks

Table 9 collects the paired effects without pooling cohorts. TCL reuses the full BPR head and seeds 42–44; its wiki RH/HR means and large seed-level spread remain in Table 2. We interpret within-seed contrasts, not native-TCL utility.

The review component check retrains history-only and representation-only heads separately (seeds 42–44). In the history-only variant, the pair loss trains the residual but has no path to the temporal encoder. The contrast between their effects measures component dependence, not mediation or the fixed-model decomposition of Appendix B.10.

InfoNCE uses separate seeds 50–52, temperature 1.0, $K \in \{1, 5\}$, three epochs and pair-loss weight 0.25. All three preregistered review directions at $K = 5$ are positive in every seed, supporting persistence beyond one-negative BPR; the small paired allocation change between the two K values does not establish a benefit from increasing K. Wiki has adverse mean pair-sampler effects at both values. These three-seed summaries establish neither significance nor general InfoNCE utility.

Protocol and sealed-result entry points. The supplement pairs the frozen acceptance-sprint protocol and Phase A summaries with the factorial; the frozen Phase B selection and acceptance protocol with seeds 45–49; the comment uniform-replay protocol with candidate replay; and the component-dependence, TCL-factorial and multi-negative protocols with their separate cohorts. The current supplement README resolves exact filenames and sealed summaries. These are compact records, not bundled raw checkpoints or per-query replay inputs.

Table 9: Paired MRR statistics for separate scope cohorts, mean ± sample SD; each cohort and estimand is specified in this subsection.

Dataset / construction	Quantity	MRR statistic
TCL / review	$RH - HR$	$+0.3760 \pm 0.0015$
TCL / wiki	$RH - HR$	-0.0813 ± 0.0378
history-only / review	$RH - RR$	$+0.2790 \pm 0.0011$
representation-only / review	$RH - RR$	$+0.0034 \pm 0.0022$
component / review	dependence contrast	$+0.2756 \pm 0.0027$
InfoNCE / review / $K = 5$	$RH - RR$	$+0.2814 \pm 0.0016$
InfoNCE / review / $K = 5$	$RH - HR$	$+0.4147 \pm 0.0021$
InfoNCE / review / $K = 5$	interaction	$+0.2241 \pm 0.0014$
InfoNCE / review	$K = 5$ minus $K = 1$ allocation	$+0.0002 \pm 0.0015$
InfoNCE / wiki / $K = 1$	$RH - RR$	-0.0381 ± 0.0286
InfoNCE / wiki / $K = 5$	$RH - RR$	-0.0252 ± 0.0222

B.10 Ranking contributions at fixed TGAT checkpoints

The recorded checkpoint decomposition complements the retrained component comparison in Section 4.5. For the strict TGAT RR and RH checkpoints, let $E = h_{\text{node}}$, $R = h_{\text{repr}}$, and $C = r_{\text{count}}(\mathbf{x}_p)$ denote the node, representation, and explicit pair-history scores. Here C uses all three history features defined in Section 2.1; it is not a raw-count-only score. Holding the checkpoint, queries, candidate set, and strict history state fixed, define

$$\Delta_C = \text{MRR}(E + R + C) - \text{MRR}(E + R). \tag{12}$$

The scores use unit inference weights; the training loss weight λ is not an inference multiplier. No model is retrained for this decomposition.

Full-model review replay. Table 4B gives the complete full-model review replay: both frozen RR/RH checkpoints, both candidate slates, all means and sample standard deviations, paired assignment effects, and historical-negative counts. Whereas that comparison evaluates the full scores under different candidate slates, the component increments below measure the effect of removing one score component from a fixed model.

Table 10: Recorded fixed-checkpoint ranking increments Δ_C for strict TGAT RR and RH, over training seeds 42–44 (mean ± sample standard deviation). Here $C = r_{\text{count}}(\mathbf{x}_p)$ uses the three history features of Section 2.1. Each increment compares scores on the same checkpoint and candidate set; it is not the effect of retraining without C. All recorded dataset/slate combinations are shown. The review uniform slate is shared across checkpoints (candidate seed 42); wiki's official slate already covers the eligible destination universe, so there is no distinct same-size uniform row.

Dataset / candidates	**Queries**	**Negatives**	**RR Δ_C**	**RH Δ_C**
tgbl-review / official	728,919	100	-0.1272 ± 0.0023	$+0.1509 \pm 0.0017$
tgbl-review / uniform	728,919	100	$+0.000050 \pm 0.000013$	-0.000974 ± 0.000004
tgbl-wiki / official	23,621	998–999	$+0.2627 \pm 0.0439$	$+0.1974 \pm 0.0661$

On review, C has opposite signed ranking increments in the RR and RH checkpoints under official candidates, while both mean increments have magnitude below 0.001 on the shared uniform slate. Thus a ranking contribution depends on the candidate set as well as the trained checkpoint. On wiki, the mean C increment is positive for both arms, although the full RH model has lower MRR than RR (Table 2). A useful component within one fixed model therefore does not imply a beneficial change in training allocation. This diagnostic neither estimates training mediation nor shows an absence of learning when an increment is small. The compact supplement retains the sealed per-seed summaries and source bindings; it omits the original checkpoints and per-query matrices.

C Diagnostic Probe and Historical-Negative Diagnostics

We first define the diagnostic probe, its defaults and feature scope, then examine gradient response, score contrast and future-positive filtering. Each diagnostic keeps its figures, controls and interpretation together; these legacy constructions do not replace the strict assignment experiment.

C.1 DTLP: A Diagnostic Probe

For direct access to node- and pair-level signals, we use DTLP (**D**iagnostic **T**emporal **L**ink **P**redictor), a lightweight additive scorer with explicitly decoupled pathways. It is a diagnostic probe, not the claim-bearing intervention. With approximately 11K parameters, DTLP permits targeted negative supervision at low computational cost. Given a temporal edge stream $\{(s_i, r_i, d_i, t_i)\}$, where r_i is the relation or event type (a single shared type on homogeneous datasets), the model maintains historical statistics with an update complexity of $O(1)$ per edge. The entity (node-level) pathway uses log in-degree and out-degree, source and destination recency, relation diversity, neighbor count, and $\log(1 + c_{\text{rel,dst}})$, where $c_{\text{rel,dst}}$ is the prior count of events sharing the relation and destination. The pair-specific pathway uses the residual's historical interaction counts (Proposition 1):

$$[\log(1 + c_{\text{triple}}(s, r, d)),\ \log(1 + c_{\text{pair}}(s, d))]\,, \tag{13}$$

Under the strict-history definition, all statistics use interactions strictly before query time t_{query}; the legacy record-order results retain the distinct history semantics described in Appendix B.2.

The optional pair-recency feature (PRF) uses Eq. 11; it distinguishes fresh from stale recurrence.

The scoring function is $\text{score}(s, r, d, t) = f_{\text{entity}}(s, r, d, t) + f_{\text{pair}}(s, r, d, t)$, where each component is a two-layer MLP. On multi-relational data, the model uses 32-dimensional relation embeddings. Because the entity and pair scorers share no parameters, the gradient magnitudes $\|\nabla_{\theta_p}\mathcal{L}\|$ and $\|\nabla_{\theta_e}\mathcal{L}\|$ of the pair-specific and entity pathways can be tracked separately.

For the probe, $\alpha \in [0, 1]$ denotes the fraction of historical hard negatives in a mixed sampler ($\alpha = 0$ random, $\alpha = 1$ all hard).

Streaming evaluation and history semantics. Training uses the first 80% of the training stream to populate statistics and the remaining 20% for gradient updates. In the streaming evaluation, model parameters remain fixed after training while pair-count statistics continue to update from preceding interactions, following the streaming approach of EdgeBank (Poursafaei et al., 2022). For the legacy record-order rows, preceding records can include interactions tied with the query timestamp; these rows are distinct from the strict timestamp-group evidence identified in Appendix B.2. Their scores characterize the recorded streaming protocols and do not establish corrected strict-history performance or show that 11K parameters suffice for general TLP. The DTLP benchmark rows should be read as streaming-probe values rather than as inductive baselines. For node-property prediction, we use DTLP-node, a separate node-property prediction extension (Table 48).

C.2 DTLP Hyperparameters

Table 11 lists the full DTLP hyperparameters.

C.3 Proposition 1 scope under the implemented feature set

Proposition 1 requires identical complete residual inputs, not zero-valued features. Legacy DTLP code offers optional PRF, cold-start and source/destination repeat-profile features alongside $c_{\text{rel,dst}}$ (Appendix C.1). Available options do not identify enabled flags; historical-row and gradient-audit bindings remain incomplete. Pair novelty fixes count features and unseen-pair recency code, not all candidate-dependent auxiliary inputs; a novel positive need not have a novel negative. Thus the novel-positive stratum alone is insufficient for Proposition 1. On tgbl-review, its average ℓ_2 pair-head gradient is 0.0160 versus the entity head's 0.1247. This stratum covers 98.3% of DTLP's post-warmup trainable edges, unlike the official-query denominator in Appendix B.4. This supports run-specific head attenuation, not exact cancellation or proved recovery.

Table 11: DTLP hyperparameters. Defaults apply unless a row specifies a dataset-specific override.

Hyperparameter	**Value**
Optimizer	Adam
Learning rate	5×10^{-3} default; 10^{-4} for yago and smallpedia
Epochs	50
Hidden dimension	32
Negative samples	10 default; 50 for yago and smallpedia; 100 for TGB-Seq rows
Warmup ratio	0.8, first 80% of training edges populate features only
Loss	BPR default; margin for yago/smallpedia; validation-selected rows may use CE/margin
Batch processing	Full-batch over trainable edges per epoch
Entity/context features	d_{out}, d_{in}, Δt_{src}, Δt_{dst}, $\lvert\mathcal{R}\rvert$, $\lvert\mathcal{N}\rvert$, $c_{\text{rel,dst}}$ (relation–destination event count; defined in Appendix C.1)
Pair-specific features	c_{triple}, c_{pair}
Relation embedding	32-dimensional learned embedding for multi-relational datasets
Training-prefix $\hat{\rho}_{\text{train}}$	Descriptive warmup estimate only; no validated policy
Selection criterion	Best validation MRR
Seeds	42, 123, 456, 789, 1024

Diagnostic questions and readouts. Table 12 separates an audited residual-input condition, a historical-pool descriptor, and two legacy training controls. These readouts have distinct settings and comparisons. They neither form one factorial nor jointly identify a ranking mechanism.

Table 12: **Questions answered by the supporting diagnostics.** The rows measure different objects under the stated settings. The strict both-novel event is a sufficient residual-input condition; the legacy overlap rate is a descriptor of the historical candidate pool. Oracle filtering and degree-profile sampling are separate legacy controls, not strict PAS treatment cells. MRR entries reproduce the reported central values for navigation; the underlying tables retain their uncertainty entries and the unresolved legacy repetition conventions.

Question	Setting and readout	Reported value	Interpretation and source
When is the audited sufficient residual-input event common?	Strict-history review; both-novel training queries (%)	99.80%	A sufficient event under Proposition 1, not all zero-gradient events. Appendix B.4, Table 1.
How often does an eligible historical candidate recur later?	Legacy wiki historical pool; future-positive overlap (%)	40.9%	A pool descriptor, not a current-time label-error rate. Appendix C.7, Table 15.
What changes under future-positive oracle filtering?	Legacy wiki, mixed-hard operating point $\alpha = 0.20$; test MRR before $\rightarrow$ after filtering	$0.624 \rightarrow 0.690$	A recorded response to an objective-changing, non-deployable oracle. Appendix C.7, Table 15.
What happens without pair-history conditioning in the degree-profile stress test?	Separate legacy wiki stress test; random versus degree-profile test MRR	0.723 vs. 0.732	Degree-profile control remains near random; realized degrees are not exactly matched. Appendix C.7, Table 16.

The overlap measurement describes which historical candidates recur, whereas the oracle and degree-profile controls describe responses to different training changes. Neither control makes overlap the uniquely identified cause of the strict assignment effect.

C.4 Gradient Diagnostics and Curriculum Controls

The historical diagnostics use two distinct measurements. Panel (a) reports ratios of minibatch-averaged total L2 gradient norms for pair and entity scorers. Panel (b) reports the pair-branch/head norm ratio G_p/G_h aggregated across minibatches, not a compositional share. The summary values and curriculum intervals retain their separate measurement identities; neither is obtained here by renormalizing the other.

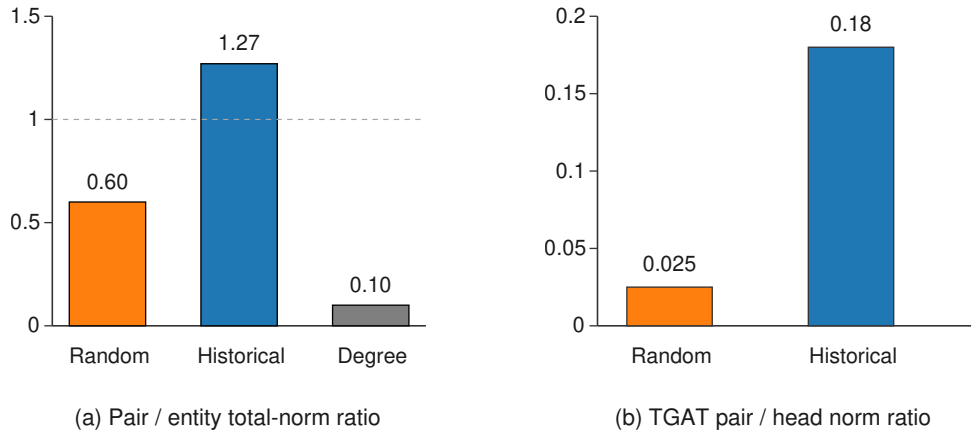


Figure 4: Historical gradient diagnostics on tgbl-review. (a) Pair/entity total-norm ratios: random 0.60, historical 1.27 and degree-based 0.10. (b) Pair-branch/head norm ratios from a separate TGAT diagnostic: random 0.025 and historical 0.18. These panels describe different measurements, not the frequency of residual cancellation.

Figure 4 retains two independent historical summaries: total pair/entity L2 norm ratios of 0.60 (random), 1.27 (historical) and 0.10 (degree-based) in the review probe, and pair-branch/head norm ratios of 0.025 and 0.18 in the separate TGAT diagnostic. The recorded wiki random-negative run ends near 0.47 for the latter ratio. None is parameter-count normalized or a compositional share. The curriculum intervals in Table 13 are a separate summary, not a rescaling of these points. These measurements characterize recorded training responses, whereas Proposition 1 concerns the residual-gradient condition for a given positive–negative pair.

Table 13: Curriculum control on tgbl-review. Pair/entity ratios use total L2 gradient norms. These intervals and the historical summaries in Figure 4 retain separate measurement identities; they are not asserted to differ only by normalization.

Negative sampler	**Final loss**	**Pair/entity gradient ratio**
Random	0.282	0.10–0.25
Historical hard	0.261	0.32–0.42
Degree-based hard	0.392	0.08–0.14

The degree-based hard sampler in this curriculum control draws uniformly from a top-total-degree pool. Figure 5 shows that these negatives raise training loss without reproducing the historical-negative pair/entity gradient response.

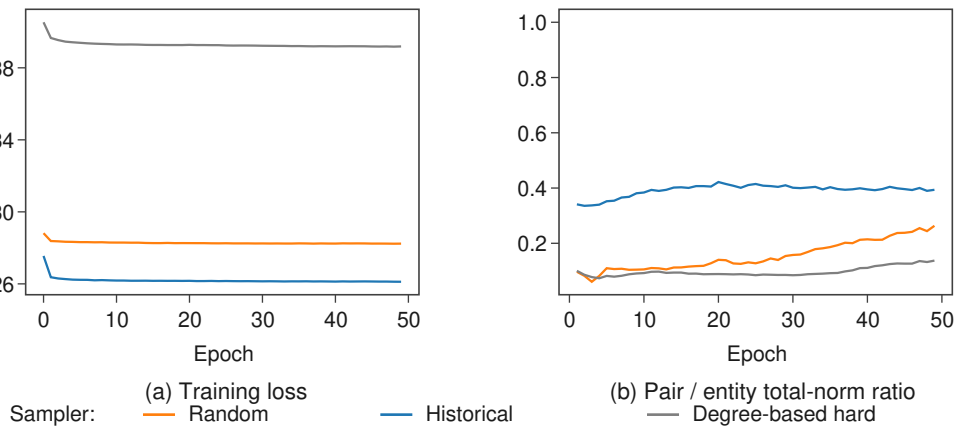


Figure 5: Curriculum control on tgbl-review. Degree-based hard negatives raise training loss but do not reproduce the historical pair-gradient response. The loss curves settle after the earliest logged updates; their original vertical scale is retained.

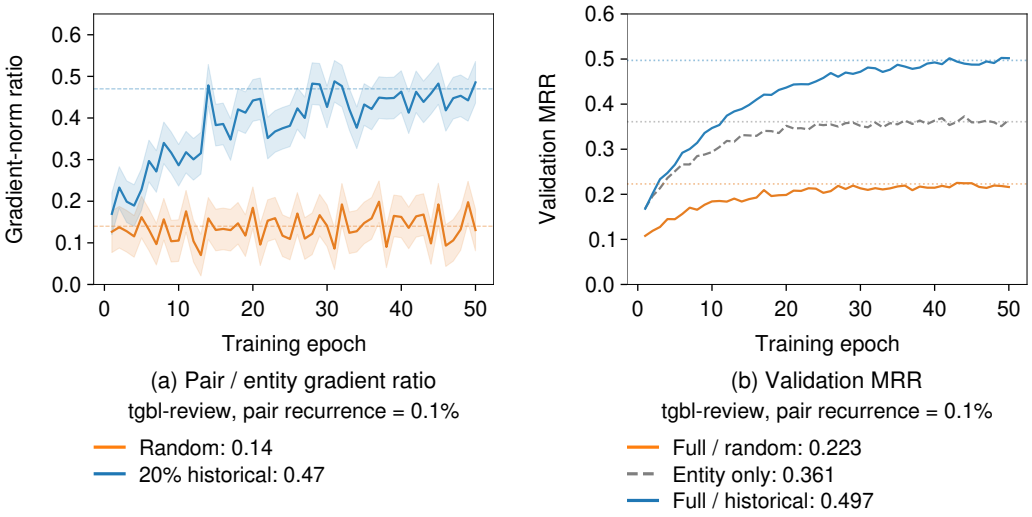


Figure 6: Pair-gradient and MRR trajectories under random and historical negative pools. Endpoint labels report within-run values. Shading retains the original reported variation; its aggregation unit is unspecified.

C.4.1 Additional Gradient Measurements

Figures 7 and 8 complement the curriculum control with measurements from the backbone and additive scorers. In Figure 7, random negatives give much smaller pair/backbone-head norm ratios than historical negatives on review; wiki retains a nonzero ratio under random negatives. Figure 8 separately reports pair and entity gradient norms and their ratio. These aggregate responses do not estimate the exact residual-cancellation event or identify mediation of ranking performance.

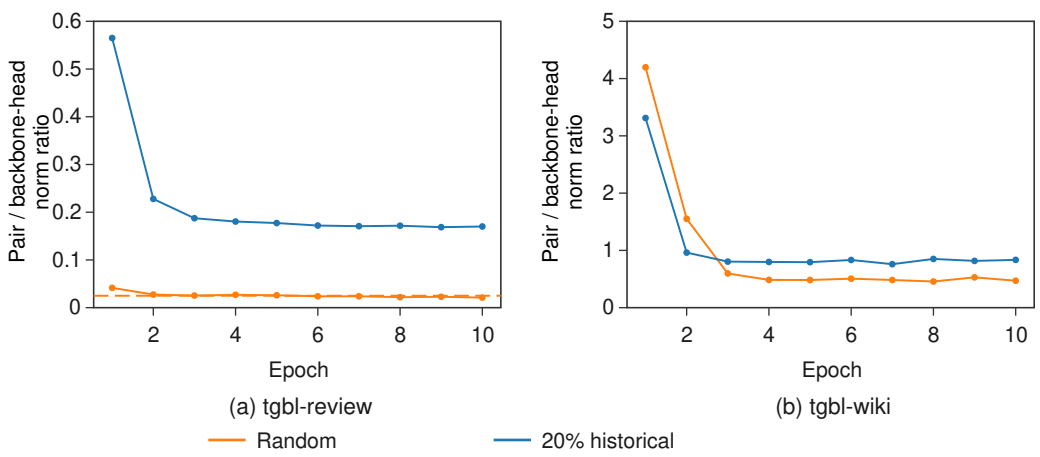


Figure 7: Backbone gradient-ratio diagnostics on tgbl-review (a) and tgbl-wiki (b). Each panel compares the recorded random-negative run with the historical-negative setting $\alpha = 0.2$, using the pair/backbone-head norm ratio shown on its axis.

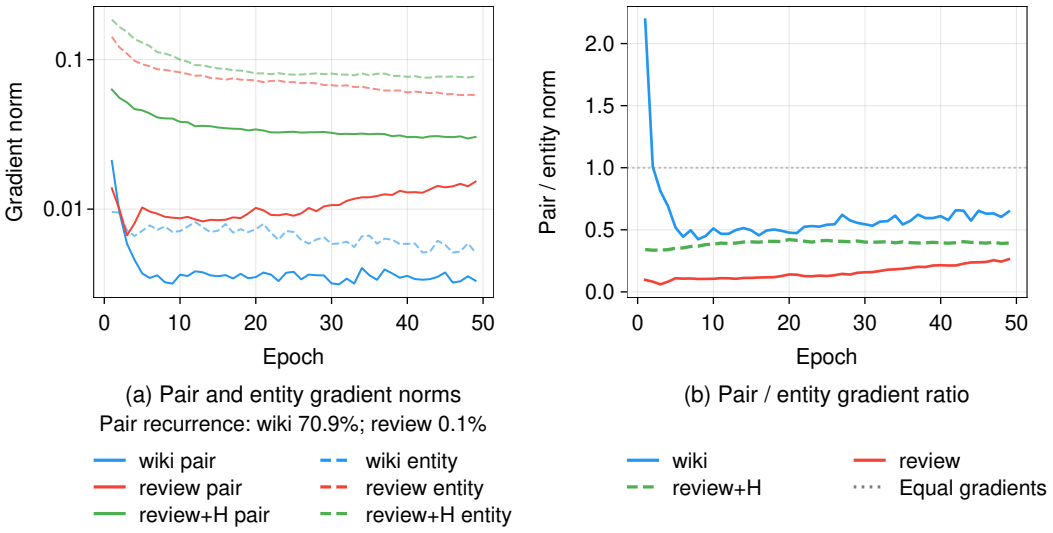


Figure 8: Gradient-norm diagnostic for the additive scorer on wiki and review. These curves supplement Figure 4; H denotes historical-negative training.

C.5 Feature and Score Contrast in Diagnostic Runs

These diagnostics separate pair-count inputs, learned pair scores and DTLP separation/MRR observations, each in its recorded setting.

Figure 10 moves from counts to learned pair scores. Its summaries describe the displayed score populations. Figure 11 separately pairs DTLP histograms with their reported separation and MRR; the two figures do not share an established checkpoint or aggregation identity.

Figure 11 pairs the original DTLP histograms with their separation/MRR annotations. Historical-negative training is accompanied by larger displayed separation and higher reported MRR in these runs.

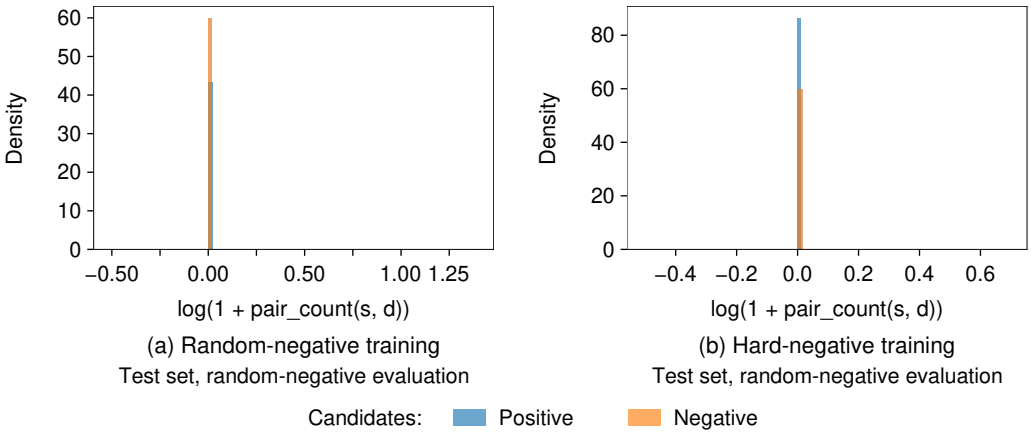


Figure 9: Pair-count feature concentration in the displayed real-data diagnostics. In both panels, 99.9% of positive and 100.0% of sampled negative features are zero. The negative-candidate legends identify random sampling in both panels. The panels retain their recorded training/evaluation labels and use linear scales. This is a feature-distribution observation, not a measurement of the full residual-input equality event.

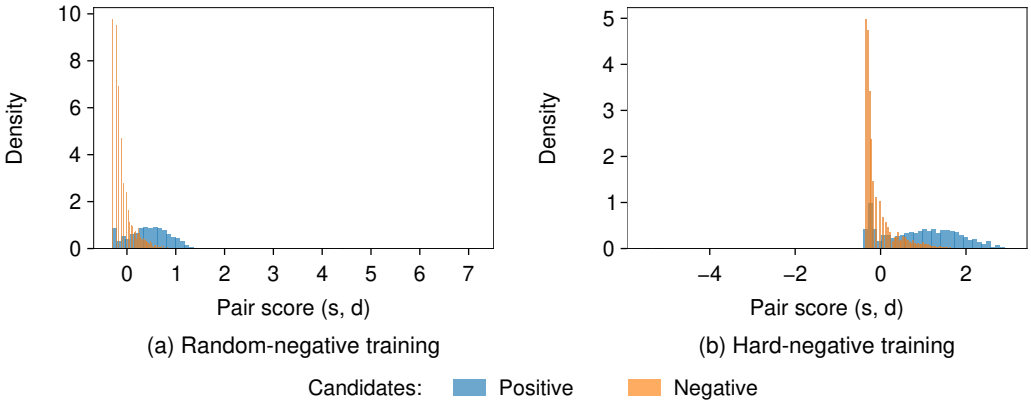


Figure 10: Real-data discrimination in the shown run. Under random training, positive/negative standard deviations are $0.4288/0.2485$ with 25.3% overlap; under hard-negative training they are $0.8713/0.4202$ with observed separation 0.9730. Panel (b) evaluates new positive pairs against random negatives after historical-negative training, and shows a visible positive–negative score contrast. Overlap and separation retain the original summary conventions, whose calculation rules are unspecified.

Figure 12(b) compares new positive pairs with historical negatives, unlike the random-negative population in Figure 10(b). Its positive-score distribution is shifted to the left of the historical-negative distribution. This separation alone does not demonstrate improved ranking by the fused scorer. The unresolved AUC scoring convention applies to Figure 12, not to the separate run-specific MRR observations in Figure 11.

Figure 13 is a synthetic high-recurrence illustration, not an additional benchmark measurement. Its highlighted 299/5,000 annotation reports future-positive overlap in that constructed population. The figure shows temporal overlap and count-feature contrast, without establishing query-time label error.

Together, these panels show why feature contrast, the direction of score separation and ranking should be examined separately. Historical training accompanies a favorable DTLP ranking observation in Figure 11; in Figure 12(b), however, new positive pairs score below historical negatives. These

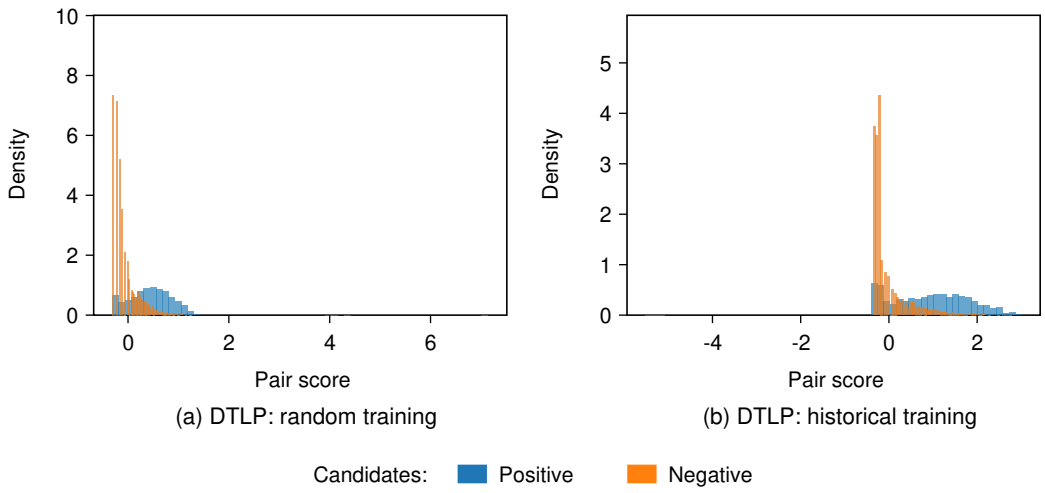


Figure 11: Recorded DTLP score-distribution diagnostics. Random-negative training (a) and historical-negative training (b) have the originally reported separation/MRR annotations of $0.506/0.214$ and $0.973/0.497$, respectively. The original histograms and axis ranges are retained. Separation follows the original, unspecified summary convention. These legacy run-specific observations are not strict factorial cells and do not identify mediation of the reported ranking difference.

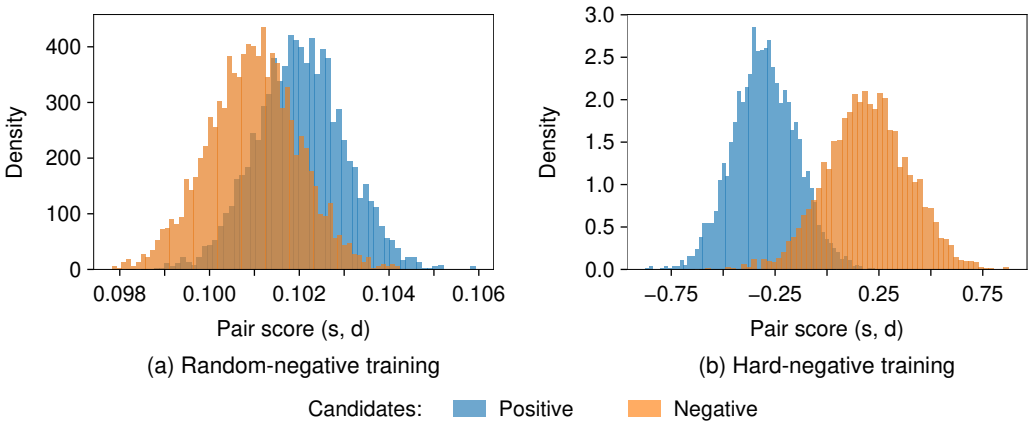


Figure 12: Pair-scorer output distributions under random-negative training (a) and historical-negative training (b). Panel (b) compares new positive pairs with historical negatives. Panel (a) has a narrow score range (standard deviation approximately 0.001). Score ranges and candidate populations differ between panels. The AUC scoring convention remains unresolved; distributional separation alone does not establish improved ranking by the fused scorer.

are observations from their respective settings; the synthetic overlap describes its own constructed population.

C.6 Sampler Effects for Discrete and Learned Pair Features

Table 14 reports the within-run interaction for discrete-count and learned pair features across review, comment, and wiki.

In these three matched diagnostic runs, both discrete count features and learned count embeddings improve with historical hard negatives on tgbl-review and tgbl-comment, while both decline on tgbl-wiki. The larger response of the discrete-count variant on tgbl-review is an empirical within-run contrast for the feature type used by the count residual, not a direct test of Proposition 1; it does

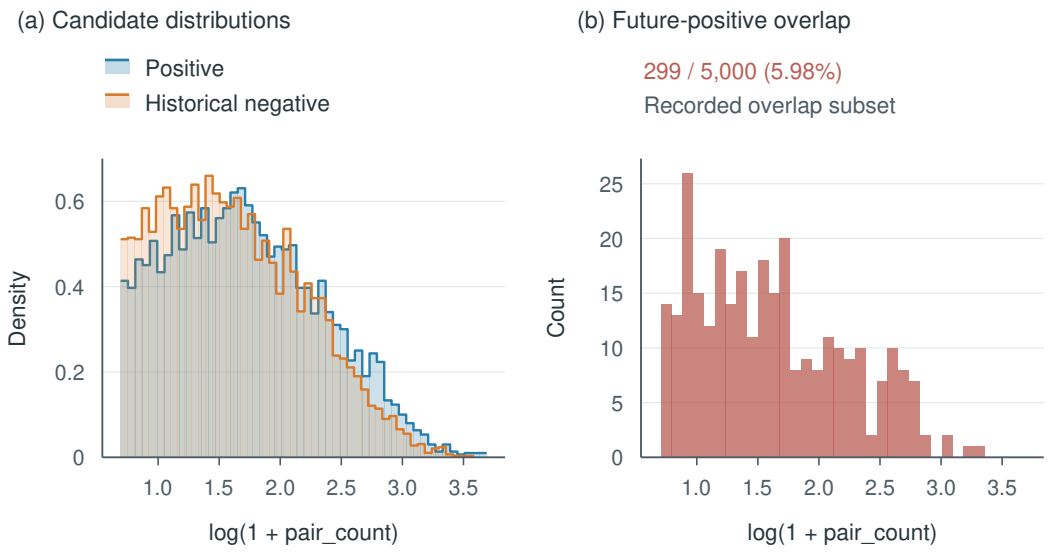


Figure 13: Synthetic high-recurrence diagnostic. (a) Count-feature densities for positive and historical-negative candidates. (b) Histogram counts for the future-positive overlap subset, shown separately from the density axis. The recorded overlap is $299/5{,}000$ (5.98%); it does not establish a query-time label error.

Table 14: Within-run feature-type interaction for discrete counts and learned embeddings. Historical hard negatives improve both variants on review/comment and reduce both on wiki; the learned-component result is empirical and outside the proposition's exact scope.

		Discrete counts			**Learned embeddings**		
Dataset	ρ	Random	Hard	Δ	Random	Hard	Δ
tgbl-review	0.001	0.172 ± 0.004	0.414 ± 0.006	$+141\%$	0.101 ± 0.003	0.181 ± 0.004	$+80\%$
tgbl-comment	0.061	0.394 ± 0.006	0.672 ± 0.007	$+71\%$	0.338 ± 0.005	0.551 ± 0.006	$+63\%$
tgbl-wiki	0.709	0.698 ± 0.001	0.679 ± 0.004	-2.7%	0.653 ± 0.004	0.637 ± 0.005	-2.5%

not establish that the proposition's cancellation mechanism mediates MRR. The parallel response of the learned variant is an empirical extension only: it neither places learned embeddings within Proposition 1 nor attributes the wiki decline uniquely to future-positive overlap.

C.7 Legacy Future-Overlap Diagnostics

Future-positive overlap also has a measurable time scale. Among 133,670 non-fallback historical draws in the recorded wiki replay, 18.88% recur within one day and 41.92% within seven days; the observed-suffix fraction is 45.61%, with a conditional median time to first recurrence of 1.50 days. A later recurrence does not change the query-time label. Appendix C.7 defines this population and its separate validation/test-window counts; no horizon-filtered training effect is inferred.

Future-positive overlap diagnostic. At query time t, let $\mathcal{F}_s(t)$ be the destinations that will appear for source s after t, and let $\mathcal{C}_s^{\text{hard}}(t) \subseteq \mathcal{H}_s(t)$ be the historical-hard-negative candidate set after the protocol's current-positive and relation/filter exclusions. Conditioned on $|\mathcal{C}_s^{\text{hard}}(t)| > 0$ and uniform sampling from $\mathcal{C}_s^{\text{hard}}(t)$, the future-positive overlap is

$$P(d^-_{\text{hard}} \in \mathcal{F}_s(t)) = \frac{|\mathcal{C}_s^{\text{hard}}(t) \cap \mathcal{F}_s(t)|}{|\mathcal{C}_s^{\text{hard}}(t)|}. \tag{14}$$

Later recurrence alone does not make a query-time negative mislabeled; a simultaneous true target is a separate case governed by the protocol's target-exclusion rule. We therefore call this quantity

future-positive overlap, not a false-negative rate. Updates on such negatives train against pairs that recur later, but overlap alone does not establish label error or harm.

Figure 14 illustrates this temporal definition.

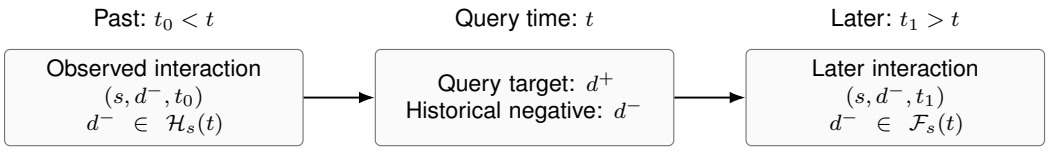


Figure 14: Schematic sequence for future-positive overlap. A historical destination sampled as a negative for the query at t appears in a later interaction. The diagram describes temporal membership; the protocol-specific rates and filtering controls are reported separately in this subsection.

These record-order diagnostics examine two distinct interventions: future-positive filtering at $\alpha = 0.20$ (Table 15) and a separate wiki comparison of random, degree-profile, and full-history negatives (Table 16). The available summaries identify seed 42 but do not specify the repetition count, repetition unit, or calculation of the reported SDs; the values are retained as descriptive results, without uncertainty-based comparisons.

Table 15: Legacy future-positive filtering at the $\alpha = 0.20$ operating point. The overlap rate is evaluated over historical-negative candidates after the recorded protocol exclusions. Filtering changes the objective; the table is a diagnostic rather than a PAS treatment cell. The reported $\pm$ values have no identified repetition unit.

Dataset	ρ	Overlap (%)	Unfiltered MRR	Filtered MRR
tgbl-wiki	0.709	40.90	0.624 ± 0.005	0.690 ± 0.006
tgbl-enron	0.978	34.40	0.509 ± 0.004	0.512 ± 0.004
tgbl-comment	0.061	2.10	0.903 ± 0.001	0.904 ± 0.001
tgbl-review	0.001	0.04	0.497 ± 0.001	0.499 ± 0.001

At the $\alpha = 0.20$ mixed-hard operating point, wiki MRR rises from 0.624 to 0.690 against the random baseline 0.721, closing 68% of that gap. The other rows show smaller recorded changes; the filtering response is setting-specific.

Table 16: Legacy wiki full-pair-history stress test, distinct from the $\alpha = 0.20$ operating point in Table 15. The degree-profile control remains near random, while future-positive filtering closes part of the historical-negative drop. This does not isolate overlap as the unique cause. The reported $\pm$ values have no identified repetition unit.

Negative mode	Conditions on pair history?	Later recurrence filtered?	Test MRR
Random	No	No	0.723 ± 0.005
Degree-profile	No	No	0.732 ± 0.005
Hard negative	Yes	No	0.250 ± 0.005
Oracle filtered	Yes	Yes	0.635 ± 0.005

In the full-pair-history stress test, filtering raises MRR from 0.250 to 0.635 against the separate 0.723 random control, closing 81% of the gap. The degree-profile control draws random destinations with probabilities proportional to their training-prefix in-degree and scores 0.732. It does not restrict draws to historical partners or require novel pairs, and realized degrees are not exactly matched. The oracle changes the objective and is non-deployable; neither control identifies overlap as the unique cause of the drop.

Table 17: Finite-horizon replay of the legacy wiki historical-negative draws. Each cell is count (rate) among 133,670 nonfallback draws. "Any later" uses the remaining stream; "evaluation future" uses the validation/test-suffix definition saved with the experiment.

Horizon	Any later	Evaluation future
1 hour	5,741 (4.2949%)	23 (0.0172%)
6 hours	12,285 (9.1905%)	282 (0.2110%)
1 day	25,233 (18.8771%)	4,518 (3.3800%)
3 days	41,235 (30.8484%)	19,682 (14.7243%)
7 days	56,038 (41.9226%)	47,782 (35.7462%)
14 days	60,964 (45.6078%)	54,896 (41.0683%)
Eventual in observed suffix	60,964 (45.6078%)	54,896 (41.0683%)

Conditional time to first qualifying interaction (days).

Definition / percentile	10	25	50	75	90	95	99
Any later	0.05	0.42	1.50	4.10	6.75	8.34	11.61
Evaluation future	1.16	2.30	3.99	5.76	7.65	9.47	11.99

C.7.1 Finite-Horizon Replay

The legacy historical-pool measurement is 40.90%; replaying recorded draws under the saved validation/test-suffix definition gives 41.07%. Table 17 separates these two future windows.

A replay of the recorded earlier wiki draws makes the temporal horizon explicit. Among 133,670 nonfallback historical draws recorded in that legacy record-order run (a different population from the official-query audit of Appendix B.4), 4.2949% recur within one hour, 18.8771% within one day, and 45.6078% at any later point of the remaining stream (41.0683% within the saved validation/test suffix); conditional median time to first recurrence is 1.50 days. Table 17 reports the complete replay of the recorded draws. These are descriptive rates, not horizon-filtered treatment effects.

C.8 Legacy count-filter diagnostic

This legacy record-order diagnostic tests a past-count filtering heuristic: replace a sampled historical hard negative (s, nd) whenever $\mathrm{paircount}(s, nd) \geq \theta \cdot \mathrm{paircount}(s, d)$ with $\mathrm{paircount}(s, nd) > 0$, using strictly past-only counts in the experiment's original record order. Because it predates the timestamp-group correction, it is contextual and does not support the strict-history claim. We evaluate it as a boundary diagnostic, not as a new headline method, testing how past-count filtering changes ranking across settings with different observed responses to historical negatives. A useful past-count proxy also requires suitable coverage and selectivity; harmful future-positive overlap alone does not guarantee that this proxy improves ranking. We compare four arms under an identical TGB-native protocol (999 negatives per query, test MRR selected by validation MRR; multi-seed: 5 seeds; evaluation uniformly capped at 4,000 queries with identical settings across arms, which reproduces the uncapped single-seed result exactly): random (no hard negatives), unfiltered hard, count-filter (θ=1), and a *non-deployable* oracle that drops a hard negative only if that candidate later appears as a positive. The three datasets are chosen by the retrospective recurrence audit (Table 52): tgbl-wiki (ρ_{diag}=70.9%; 40.9% of its historical hard negatives later appear as positives), tgbl-subreddit (86.5%, hard negatives near-inert), and tgbl-lastfm (87.1%, a second documented exception in Table 52: aggregate ρ high, yet hard negatives mildly *help*, $+0.8\%$).

Table 18 reports the five-seed legacy record-order count-filter diagnostic and its accounting fields.

Three observations follow within this earlier experimental setup. First, the unfiltered arms reproduce the main table's direction on tgbl-wiki and its near-null magnitude on tgbl-subreddit and tgbl-lastfm: hard negatives cost 7.2 MRR points on tgbl-wiki (main table -13%); the subreddit sign differs within noise ($+0.1\%$ here versus -0.6% there); and lastfm is $+0.8\%$. Second, on tgbl-wiki the filter recovers 6.8 of the 7.2 lost points ($0.646 \rightarrow 0.714$), ending within 0.4 points of random and *above* the later-outcome oracle filter (0.683). The proxy and oracle remove different candidate sets because the proxy also fires on high-count candidates that do not recur later; this diagnostic does not isolate why

Table 18: Legacy record-order count-filter diagnostic: test MRR (mean±sd, 5 seeds). ρ_{diag} is the retrospective diagnostic (Table 52). Filtered is the filter firing rate; later recurring among filtered is the fraction of filtered negatives that recur later, reported descriptively and never used by the filter.

Dataset	ρ_{diag}	Random	Hard	Count-filter	Oracle	Filtered (%)	Later recurring among filtered (%)
tgbl-wiki	70.9%	0.718 ± 0.006	0.646 ± 0.008	0.714 ± 0.006	0.683 ± 0.006	38.3	44.9
tgbl-subreddit	86.5%	0.757 ± 0.001	0.758 ± 0.001	0.759 ± 0.000	0.757 ± 0.001	29.6	53.1
tgbl-lastfm	87.1%	0.095 ± 0.001	0.096 ± 0.001	0.092 ± 0.001	0.094 ± 0.001	38.7	58.2

the resulting MRRs differ. A single-seed θ-sweep on wiki is monotone, and every tested $\theta \in [0.5, 3]$ beats unfiltered hard. Third, on borderline tgbl-lastfm *both* filters mildly hurt—count-filter -0.4 and even the oracle -0.2 points versus unfiltered hard—even though half of the sampled hard negatives there later appear as positives (oracle filter rate 50.9%). This agrees with the descriptor audit of Appendix G.4, from the measurement side (tgbl-lastfm helps at c_d=0.509): a high forward-overlap *rate* does not by itself imply that hard negatives are net-negative on this dataset; filtering either later-positive candidates or the count-proxy set lowers MRR in this experimental setup. The experiment does not resolve why this high-overlap row has that filtering response; the descriptive exception is recorded in Table 52. The filter experiment does not identify a unique mechanism. We therefore make no deployment recommendation from this diagnostic and do not treat it as instantiating the strict PAS mechanism.

D Diagnostic-Probe Ablations and Sensitivity

The ablations below use the diagnostic probe defined in Appendix C.1, retaining their own dataset, candidate, loss and seed grids. The historical-sample fraction α differs from strict pathway assignment; fixed-statistic evaluation tests streaming dependence, not strictness or inductivity.

D.1 DTLP as a Recurrence Stress Test

Table 19 tests the streaming probe across recurrence regimes, retaining each row's feature and candidate protocol.

Table 19: DTLP stress-test values, not a common leaderboard. Descriptive ρ uses test/pretest overlap for review/comment, a warmup estimate for mooc, and source-specific legacy windows elsewhere. Bold marks protocol-matched comparisons only. Mooc uses 20 negatives versus CRAFT-R's 100 and supports no direct ranking. ‡Yago uses the legacy 40K-query pair-only margin-loss run, matched to our Repeat/Recurrency comparator. The dictionary-based yago result and reference-method definitions are given in Appendix F.2; full protocols are in Appendix F.

Dataset	ρ (%)	DTLP MRR	Reference method	Reference MRR
tgbl-review	0.1	0.497 ± 0.001	GraphMixer	0.521 ± 0.015
tgbl-comment	6.1	0.919 ± 0.001	CRAFT	0.917 ± 0.006
thgl-github	12.1	0.818 ± 0.007	EdgeBank	0.413
thgl-software	23.1	0.547 ± 0.003	STHN	0.731 ± 0.005
mooc	59.8	0.596 ± 0.022	CRAFT-R	0.623 ± 0.004
tkgl-yago	74.3	0.875 ± 0.018	Repeat/Recurrency‡	0.803

The probe trails GraphMixer on review and STHN on software, while the matched github and legacy yago comparisons are favorable. These mixed results bound what the diagnostic features capture; mooc supports no cross-method ranking because its candidate counts differ.

Fixed-statistic test-time evaluation drops DTLP MRR by 22–39% across selected datasets (Appendix D.7); this measures dependence on test-stream statistic updates, not timestamp strictness.

D.2 Feature and Sampler Sensitivity in DTLP

Table 20 examines feature-pathway removal and the hard-negative fraction within the same diagnostic probe.

Table 20: DTLP ablations. "Full" is the random-negative probe ($\alpha = 0$), not the strict TGAT RR condition in Table 2. Top: feature-pathway ablations at $\alpha = 0$, with a separate Hard-neg row at $\alpha = 0.20$; bottom: the α sweep. Bold marks the best value in each column within each block.

Config	**tgbl-wiki** $\rho = 70.9\%$	**tgbl-review** $\rho = 0.1\%$	**tgbl-comment** $\rho = 6.1\%$
Feature-pathway ablations ($\alpha = 0$); Hard-neg shown separately			
Full	0.721 ± 0.003	0.223 ± 0.009	0.675 ± 0.010
Entity-only	0.060 ± 0.001	0.361 ± 0.001	0.744 ± 0.002
Pair-only	0.698 ± 0.001	0.175 ± 0.003	0.392 ± 0.005
Hard-neg ($\alpha = 0.20$)	0.624 ± 0.005	0.497 ± 0.001	0.903 ± 0.001
α sweep on the same probe			
$\alpha = 0$ *(=Full)*	0.721 ± 0.003	0.223 ± 0.009	0.675 ± 0.010
$\alpha = 0.05$	0.715 ± 0.004	0.501 ± 0.003	0.882 ± 0.006
$\alpha = 0.10$	0.688 ± 0.006	0.496 ± 0.007	0.897 ± 0.004
$\alpha = 0.20$	0.624 ± 0.005	0.497 ± 0.001	0.903 ± 0.001
$\alpha = 0.20$ + PRF	0.686 ± 0.008	0.500 ± 0.003	0.919 ± 0.001
$\alpha = 0.50$	0.449 ± 0.007	0.492 ± 0.005	0.898 ± 0.006

Interpretation. Under random negatives on review, the full probe scores below the entity-only probe (0.223 vs. 0.361), while pair-only scores 0.175. On wiki, pair-only nearly matches full (0.698 vs. 0.721); at $\alpha = 0.20$ the diagnostic probe improves on review/comment and degrades on wiki, with PRF separating fresh from stale recurrences. These are selected within-probe dose-response curves, not a fitted recurrence rule. The review change ($0.223 \rightarrow 0.497$) shares the sign of strict TGAT $RH - RR$ ($0.2046 \rightarrow 0.4855$; Table 2) but remains descriptive because the models and interventions differ; it is not a replication.

Table 40 gives a separate native-TGB recency check: adding PRF raises tgbl-comment MRR from 0.903 to 0.919 under hard negatives and tgbl-wiki from 0.721 to 0.744 under random negatives.

D.2.1 Historical-Negative Ratio

The lower block of Table 20 varies the historical negative fraction in the same DTLP probe. Every tested nonzero α setting improves review and comment relative to $\alpha = 0$, whereas wiki declines across the five sampler-ratio settings. The $\alpha = 0.20 +$ PRF row additionally changes the feature set and is not an extra point on the pure sampler-ratio curve. Figure 15 visualizes the wiki and review means and reported standard deviations; the comment values remain in the complete table. These probe-specific responses neither repeat the strict factorial nor fit a recurrence rule.

D.2.2 Additional Sampler-Ratio Checks

The main α sweep (Table 20) covers three TLP datasets. We extend it to two TKG datasets and the near-threshold THG dataset thgl-github ($\rho = 12.1\%$). These seed-42, CRAFT-R-protocol sweeps are descriptive dose-response checks, not a validation of a recurrence rule.

Table 21 extends the descriptive α sweep to the two TKG rows and near-threshold thgl-github under the captioned protocols.

The observed TKG curves are inverted-U shaped, while thgl-github drops by 5.7% at $\alpha = 0.5$. Because the rows differ from the main strict factorial and use one seed, we do not infer a recurrence-determined transition from these shapes.

Two-seed checks on tkgl-yago (seeds 42 and 43) give 0.832 ± 0.002 at α=0 and 0.753 ± 0.060 at α=0.5; the latter dispersion is comparable to the mean α shift, so they do not establish broad stability.

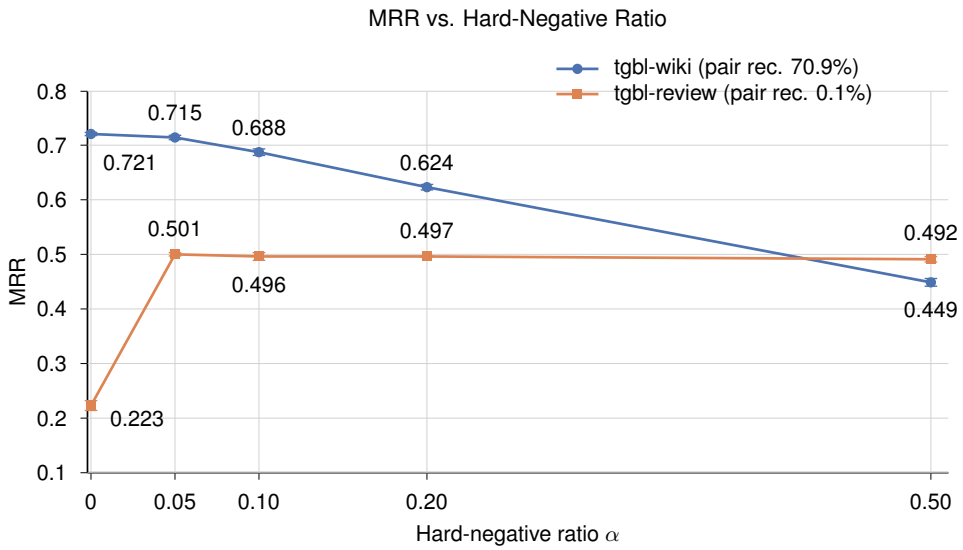


Figure 15: Hard-negative ratio sweep on wiki and review using all five α settings from Table 20. Markers show the table's means; error bars show the corresponding reported standard deviations.

Table 21: Extended descriptive α sweep. The two TKG rows show an inverted-U in these runs; thgl-github is near-threshold ($\rho = 12.1\%$), not low recurrence. The $\alpha = 0$ anchors use the BPR-style probe under the CRAFT-R protocol and are not comparable to margin-loss or TGB-native cells; seed 42 unless noted. An em dash means that the intermediate α settings were not evaluated for thgl-github. Bold marks the best value in each row.

Dataset	**Regime**	**α=0**	0.05	0.10	0.20	0.50
tkgl-smallpedia	mid-ρ TKG	0.637	0.773	**0.779**	0.776	0.603
tkgl-yago	mid-ρ TKG	0.833	0.851	**0.853**	0.838	0.710
thgl-github	near-threshold THG	**0.795**	—	—	—	0.750

D.3 Ablation Study Across Datasets

Tables 22–23 report ablations across datasets.

Table 22: Ablation study across datasets under the default feature-ablation grid. Pair-only versus entity-only contributions vary by dataset within this grid and do not validate a recurrence threshold; tuned margin-loss TKG results are reported separately in Table 42. For tkgl-yago, the $0.494 \rightarrow 0.192$ BPR pair is the matched default-grid intervention contrast in Table 52.

Config	**thgl-software**	**tkgl-smallpedia**	**tkgl-yago**	**tgbl-wiki**	**tgbl-review**	**tgbl-comment**
Full model	0.547 ± 0.003	0.330 ± 0.006	0.494 ± 0.006	0.721 ± 0.003	0.223 ± 0.009	0.675 ± 0.010
+ hard-neg	0.516 ± 0.006	0.321 ± 0.005	0.192 ± 0.007	0.624 ± 0.005	0.497 ± 0.001	0.903 ± 0.001
Pair only	0.498 ± 0.005	0.509 ± 0.008	0.334 ± 0.007	0.698 ± 0.001	0.175 ± 0.003	0.392 ± 0.005
Entity only	0.113 ± 0.004	0.006 ± 0.001	0.005 ± 0.003	0.060 ± 0.001	0.361 ± 0.001	0.744 ± 0.002
ρ	23.1%	25.0%	74.3%	70.9%	0.1%	6.1%

Within this default feature-ablation grid, pair-only dominates entity-only on the TKG/THG rows, whereas entity-only exceeds the full random-negative probe on review/comment. The matched yago BPR hard-negative decline belongs to this grid; the tuned margin-loss result remains a separate benchmark cell.

D.3.1 Heuristic Feature Omission

The separate additive-heuristic omission check in Figure 16 is not an ablation of trained BPR/margin-loss DTLP or PAS. Its signed bars describe whether removing each heuristic feature raises or lowers MRR.

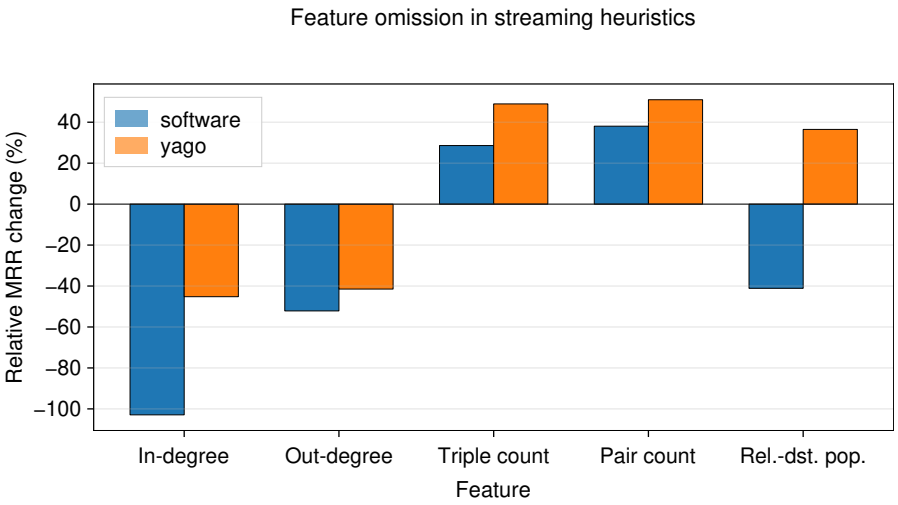


Figure 16: Heuristic feature omission (3,000 queries per dataset). Bars: $100(\mathrm{MRR}_{\mathrm{all}} - \mathrm{MRR}_{\mathrm{removed}})/\mathrm{MRR}_{\mathrm{all}}$. Positive means removal lowers MRR; negative means it raises MRR.

D.4 Additional Entity- and Pair-Only Ablations

Table 23: Entity-only versus pair-only ablation across TGB-native datasets, including the primary regime-spanning rows; rows use the native no-collision-check protocol unless noted.

Dataset	ρ (%)	Full	Entity-only	Pair-only	Dominant
tgbl-review	0.1	0.223 ± 0.009	0.361 ± 0.001	0.175 ± 0.003	Entity
tgbl-comment	6.1	0.675 ± 0.010	0.744 ± 0.002	0.392 ± 0.005	Entity
tgbl-lastfm	87.1	0.120 ± 0.001	0.047 ± 0.003	0.118 ± 0.004	Pair
tgbl-coin	85.8	0.773 ± 0.005	0.725 ± 0.004	0.724 ± 0.006	Balanced
tgbl-subreddit	86.5	0.737 ± 0.001	0.193 ± 0.004	0.734 ± 0.003	Pair
tgbl-enron	97.8	0.516 ± 0.004	0.140 ± 0.004	0.443 ± 0.006	Pair
tgbl-uci	65.7	0.326 ± 0.003	0.109 ± 0.004	0.321 ± 0.004	Pair
tgbl-wiki	70.9	0.721 ± 0.003	0.060 ± 0.001	0.698 ± 0.001	Pair

In this ablation grid (Table 23), entity-only exceeds the full random-negative model on review and comment, while pair-only nearly matches full on several high-recurrence rows. This is a descriptive within-grid pattern, not a recurrence-threshold result. The balanced tgbl-coin row shows that recurrence alone is insufficient to determine pathway dominance.

D.5 Hard-Negative Impact on Additional Datasets

Table 24 reports additional within-row effects. The single-seed tgbl-flight cell improves despite a destination set of 18,143 nodes; together with coin, it cautions against treating recurrence alone as a decision rule. We do not infer its per-query overlap mechanism from graph size.

Table 24: Hard-negative impact on additional TGB-native datasets with no collision check. Cells with uncertainty show mean±std; the tgbl-flight hard-negative cell is a single-run point estimate. Deltas are computed from full-precision values. †tgbl-flight uses margin loss with PRF; baseline is the matched DTLP + PRF random-negative row from Table 38, and the hard-neg cell uses $\alpha = 0.20$. The tgbl-flight ρ cell is omitted: that 86.7% DyGLib Flights value was not remeasured on tgbl-flight.

Dataset	ρ (%)	Baseline	+ hard-neg	Δ	Runs
tgbl-enron	97.8	0.516 ± 0.004	0.500 ± 0.005	-3.1%	3
tgbl-lastfm	87.1	0.120 ± 0.001	0.121 ± 0.001	$+0.8\%$	3
tgbl-uci	65.7	0.326 ± 0.003	0.215 ± 0.008	-34.0%	3
tgbl-subreddit	86.5	0.737 ± 0.001	0.732 ± 0.001	-0.6%	3
tgbl-coin	85.8	0.773 ± 0.005	0.808 ± 0.005	$+4.5\%$	3
tgbl-flight†	–	0.885 ± 0.003	0.901	$+1.8\%$	1

D.6 Recorded Results by Training Loss

Table 25 groups the recorded probe results by training objective. The reported MRRs are close on forum (0.802–0.804) and span a wider range on yago (0.494–0.916). These are observations across recorded configurations: the available records do not establish matched evaluation candidates for the unmarked rows, so the differences cannot be attributed to the loss alone. Table 45 provides the separate full-destination DyGLib comparison.

Proposition 1 covers a separately parameterized residual under pairwise-difference objectives, including BPR and score-difference margin. The CE rows and other components do not extend that exact conclusion.

Table 25: Recorded diagnostic-probe MRR by training objective. The unmarked rows describe configurations whose evaluation-candidate protocols are not documented sufficiently to support a matched loss comparison. ‡The separate legacy 40K-query tkgl-yago row uses the pair-only margin-loss protocol; an em dash means that BPR or CE was not evaluated under that protocol.

Dataset	BPR	CE	Margin
tgbl-enron	0.516 ± 0.004	0.521 ± 0.004	0.522 ± 0.005
tgbl-lastfm	0.120 ± 0.001	0.123 ± 0.003	0.121 ± 0.004
tgbl-subreddit	0.737 ± 0.001	0.730 ± 0.004	0.728 ± 0.006
tgbl-uci	0.326 ± 0.003	0.338 ± 0.004	0.338 ± 0.004
thgl-forum	0.802 ± 0.004	0.804 ± 0.005	0.803 ± 0.004
tgbl-wiki	0.721 ± 0.003	0.740 ± 0.004	0.741 ± 0.005
tkgl-yago	0.494 ± 0.006	0.860 ± 0.005	0.916 ± 0.015
tkgl-yago (legacy 40K)‡	—	—	0.875 ± 0.018

D.7 Fixed-Statistic Evaluation of DTLP

DTLP also exposes dependence on streaming statistics. With its trained parameters fixed, freezing the test-time statistics lowers MRR by 22–39% on the four reported datasets (Appendix D.7). This comparison measures the value of statistic updates in those runs, not timestamp strictness or general inductive capacity. DTLP's definition, gradient controls, and streaming ablations are collected in Appendices C–D; their protocols are distinct from the strict factorial.

To quantify DTLP's dependence on test-stream updates to pair-count and entity statistics, we re-evaluate the same trained models under a fixed-statistic test protocol (Table 26): the streaming statistics recorded at the end of validation are held fixed, and test edges do not update these statistics.

The streaming-vs-fixed gap ranges 22–39% across these four selected datasets. DTLP should therefore be read as a *streaming probe* (in the spirit of EdgeBank), not as an inductive parameter-count-controlled comparator. This control freezes DTLP's engineered statistics. It does not measure the effect of freezing the histories used by DyGFormer, TGAT, or TPNet; fixed model parameters

Table 26: DTLP MRR under streaming vs. fixed-statistic test-time evaluation. Streaming numbers are taken from the main TGB-LP / TGB-native tables; fixed-statistic numbers hold the streaming statistics recorded at the end of validation fixed.

Dataset	ρ **(%)**	**Streaming**	**Fixed-statistic**
tgbl-review	0.1	0.497 ± 0.001	0.384 ± 0.003
tgbl-wiki	70.9	0.721 ± 0.003	0.563 ± 0.005
tgbl-comment	6.1	0.903 ± 0.001	0.553 ± 0.004
tgbl-coin	85.8	0.808 ± 0.005	0.575 ± 0.006

do not by themselves imply a fixed interaction history. Streaming numbers in main tables are the intended streaming-probe evaluation; fixed-statistic numbers quantify dependence on test-stream statistic updates and are not a substitute for the main-table values, nor evidence of protocol-wide robustness or recurrence independence.

D.7.1 Streaming-Heuristic Trajectories

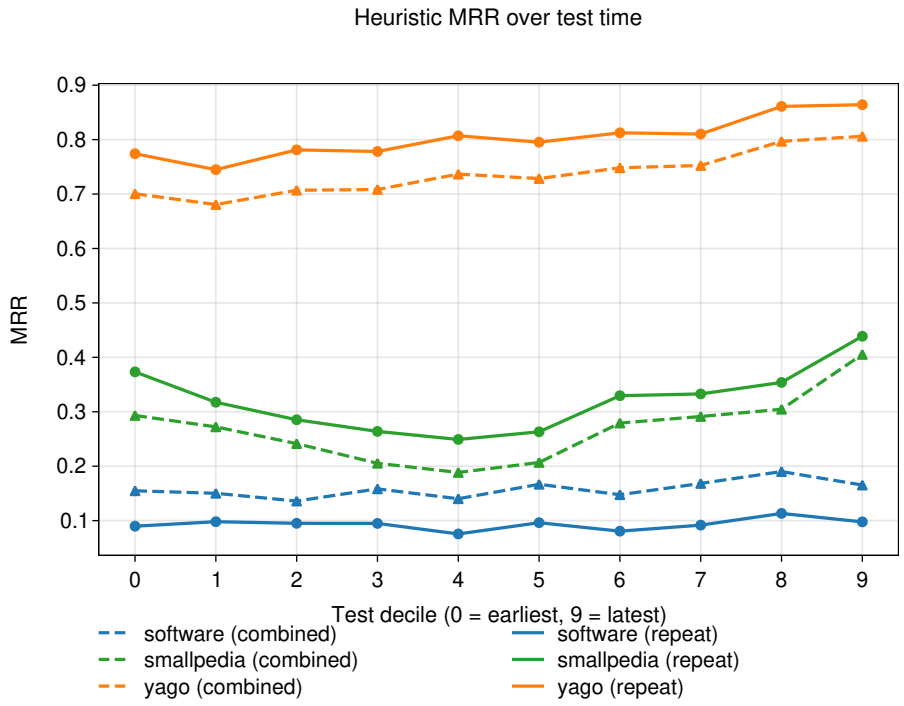


Figure 17: MRR of the displayed streaming heuristics over test deciles. The trajectories include declines and recoveries; the figure has no comparison that holds test-time statistics fixed.

The decile trajectories fluctuate through declines and recoveries. Unlike the fixed-statistic control above, this figure has no counterfactual that holds statistics fixed; it describes changes over the test stream rather than identifying the effect of statistic updates.

D.8 Multi-Seed Variance

For representative configurations, we report five random seeds (42, 123, 456, 789, and 1024). Table 27 lists means, standard deviations, and individual seed results. Standard deviations are computed from the full-precision runs; individual seed results are rounded to three decimal places for auditability.

Table 27: Multi-seed means, standard deviations, and individual values for representative configurations.

Dataset / Configuration	Mean	Std	Individual Seeds
tgbl-wiki full	0.721	0.003	0.724, 0.725, 0.720, 0.718, 0.717
tgbl-wiki pair-only	0.698	0.001	0.697, 0.697, 0.699, 0.698, 0.699
tgbl-review hard-neg	0.497	0.001	0.496, 0.497, 0.496, 0.498, 0.498
tgbl-review entity-only	0.361	0.001	0.362, 0.359, 0.360, 0.362, 0.362
tgbl-review full	0.223	0.009	0.214, 0.218, 0.231, 0.220, 0.234
tgbl-comment hard-neg	0.903	0.001	0.904, 0.903, 0.902, 0.904, 0.904
tgbl-comment full	0.675	0.010	0.675, 0.681, 0.670, 0.661, 0.687
TGAT + pair-residual review	0.515	0.001	0.515, 0.516, 0.516, 0.515, 0.515
DyGFormer + pair-residual wiki	0.798	0.004	0.790, 0.800, 0.799, 0.800, 0.801

For the five paired seeds shared by the entity-only and pair-only diagnostics, pair-only minus entity-only test MRR is -0.1862 ± 0.0027 on review and $+0.6378 \pm 0.0011$ on wiki. Two-sided paired-t tests remain significant after Holm correction ($p = 1.08 \times 10^{-8}$ and 4.80×10^{-12}, respectively), and every paired difference has the reported sign. With only five pairs, however, the exact two-sided sign-test value is $p = 0.0625$ for each dataset; comment is excluded because a complete paired source is not available. These post-hoc tests support the mean contrasts under their stated assumptions, not a broader population-level reversal.

D.8.1 Independent consecutive-seed checks for selected rows

Single-seed test MRRs for seeds 43–46 (default configurations of the paper):

- tkgl-yago: seed 43 = 0.813, seed 44 = 0.820, seed 45 = 0.850, seed 46 = 0.840. **5-seed mean 0.830 $\pm$ 0.014** vs. the seed-42 probe baseline 0.827 of Table 28.
- thgl-software: seed 43 = 0.531, seed 44 = 0.528, seed 45 = 0.537, seed 46 = 0.533. **5-seed mean 0.535 $\pm$ 0.008** vs. the main-table value 0.547.
- thgl-github: seed 43 = 0.808, seed 44 = 0.803, seed 45 = 0.802, seed 46 = 0.800. **5-seed mean 0.806 $\pm$ 0.006** vs. the main-table value 0.818.
- tgbl-coin (TPNet+PAS): seed 43 = 0.9464, seed 44 = 0.9441, seed 45 = 0.9443, seed 46 = 0.9426. **5-seed mean 0.943 $\pm$ 0.003** compared with the separately reported TPNet+PAS coin cell 0.939 (Table 38).
- tgbl-comment (TPNet+PAS): seed 43 = 0.9403, seed 44 = 0.9441, seed 45 = 0.9405, seed 46 = 0.9428. **5-seed mean 0.943 $\pm$ 0.002** compared with the separately reported TPNet+PAS comment cell 0.945 (Table 38).

Main-table means use the declared seeds 42/123/456/789/1024, whereas this independent audit combines seed 42 with consecutive seeds 43–46. They are distinct reruns; uncertainty from one seed set is not attached to the other's mean.

D.9 Reverse-edge feature ablation across TKG/THG datasets

We mask the reverse-edge count feature (set to zero) and re-evaluate DTLP on four TKG/THG datasets. Across these four single-seed rows, the reported relative changes under reverse-edge-count masking are no lower than -3.6%; this does not establish universal independence from reverse-edge features. The earlier preliminary -71% figure on thgl-software was traced to an unrelated interaction with the relation-aware-pair configuration that simultaneously zeroed the forward pair count; the isolated reverse mask gives -3.1%.

Table 28 reports the seed-42 reverse-edge-count-mask ablation for the listed TKG/THG rows.

Table 28: Reverse-edge-count mask ablation under seed 42. For thgl-software/myket/github, the unmasked value matches the corresponding main-table cell at reported precision. The tkgl-yago value is the seed-42 probe baseline for this ablation protocol, not a main-table cell. The masked column changes only the reverse-edge count feature.

Dataset	**Unmasked seed-42 MRR**	**Reverse-masked MRR**	**Δ**
tkgl-yago	0.827	0.7976	-3.6%
thgl-software	0.547	0.5302	-3.1%
thgl-myket	0.904	0.9090	$+0.6\%$
thgl-github	0.818	0.8036	-1.8%

D.10 Inverse-relation leakage audit on TKG benchmarks

We audit whether tkgl-yago or tkgl-smallpedia contain explicit inverse-relation pairs (r, r') such that (s, r, d, t) test edges can be predicted from (d, r', s, t') training prefix edges with $t' < t$. For each ordered relation pair we measure the fraction of (s, r, d) edges whose reciprocal (d, r', s) appears within 1 day. On tkgl-yago, zero relation pairs pass a 50% threshold; on tkgl-smallpedia four candidate pairs pass, but the weighted fraction of test edges with an inverse-counterpart in the training prefix is only **0.43%**. Under this criterion, the measured coverage is too small to provide a dataset-wide explanation for the reported yago and smallpedia results.

We also audited tkgl-polecat and tkgl-wikidata: 0 ordered relation pairs passed the 50% symmetry threshold on tkgl-polecat (benchmark-wide weighted leak 0.00%), and 3 pairs passed on tkgl-wikidata (benchmark-wide weighted leak 0.0006%, 9 of 1,478,431 test edges). The tkgl-wikidata pairs are extremely rare (only 10 covered test edges, 9 with prefix reciprocals), so the audit still indicates negligible benchmark-level inverse-relation leakage across all four TKG benchmarks.

E Contextual Architecture Constructions

We distinguish plug-in compatibility, matched PDRE reruns and external CRAFT retrofits. Each construction has its own architecture, fixed-backbone rule, evaluation protocol and local controls.

E.1 Training allocation in the historical submitted construction

Table 29 preserves the TGAT dual-path comparison in the earlier submitted manuscript. This historical record-order construction includes a stop-gradient variant; values and $\pm$ are retained as originally reported, without assigning current strict-factorial seeds, loss weights or timestamp semantics to it. The submitted comparison lacks HR and is distinct from the complete four-arm test in Table 2.

Table 29: MRR in the historical submitted construction, retaining the original values and reported uncertainty.

Training mode	tgbl-review	tgbl-comment	tgbl-wiki
Joint random	0.328 ± 0.018	0.670 ± 0.005	0.788 ± 0.005
Joint historical	0.132 ± 0.005	0.573 ± 0.009	0.636 ± 0.006
Decoupled, no stop-gradient	0.479 ± 0.006	0.832 ± 0.004	0.660 ± 0.006
Historical PAS (stop-gradient)	0.474 ± 0.007	0.853 ± 0.004	0.661 ± 0.006

E.2 PDRE Recipe and Construction-specific Settings

Algorithm 1 describes the PDRE variant whose backbone is held fixed, not the joint PAS procedure in Section 2. Historical hard negatives are current-positive-excluded pair-history candidates used for training, without oracle future-positive filtering unless a table caption states otherwise; this is separate from evaluation-time collision checking.

Algorithm 1 Pathway-Decoupled Residual Expert (PDRE): three-phase residual pair-expert training and inference

1: Split training edges chronologically into T_{base}, T_{res}, T_{gate}, and V_{gate}.
2: Train backbone f_θ on T_{base} with random negatives; hold θ fixed.
3: Cache pair features causally and train residual expert r_ψ on T_{res} with current-positive-excluded historical hard negatives; hold ψ fixed.
4: Train gate g_ϕ on T_{gate} with L1 sparsity using BPR on the fused score $\text{score} = f_\theta + g_\phi r_\psi$
5: At inference, $\text{score}(s,d,t) = f_\theta(s,d,t) + g_\phi(s,d,t)\, r_\psi(s,d,t)$.

The gate input combines normalized backbone score, residual score, log pair count, and recency:

$$\mathbf{x}_{\text{gate}} = (\tilde{\ell}_B,\ r,\ \log(1 + \widetilde{c_{\text{pair}}}(s,d)),\ \text{recency}(s,d,t)). \tag{15}$$

A two-layer MLP with hidden dimension 16 outputs $g(s,d,t) = \sigma(g_\phi(\mathbf{x}_{\text{gate}})) \in [0,1]$. The gate loss is

$$\mathcal{L}_{\text{gate}} = \frac{1}{K}\sum_{k=1}^{K} -\log\sigma(\text{score}(s,d^+,t) - \text{score}(s,d_k^-,t)) + \lambda_{\text{L1}}|g(s,d^+,t)|. \tag{16}$$

We use AdamW with learning rate 10^{-3}, weight decay 10^{-4}, up to 20 epochs, early stopping with patience 5, and gate bias initialized as $b_0 = \text{logit}(\max(0.01,\ 1 - 5\hat{\rho}_{\text{train}}))$. This $\hat{\rho}_{\text{train}}$-dependent initialization is construction-specific to the PDRE gate and is not a validated deployment selector or a PAS requirement. Table 30 distinguishes PDRE from strict PAS and contextual variants.

Table 30: Separate settings for PAS and contextual variants, including PDRE. Strict PAS uses branch-wise BPR losses and fixed weights without a learned gate (Section 2); stop-gradient and learned-gate entries are separate variants.

Component	Configuration
Strict PAS: full additive main factorial	
Training	Branch-wise BPR; $L_N^X + 0.25L_P^Y$; three epochs; joint encoder updates
Inference	Unit-weight $s_N + s_P$; no learned gate
PDRE: held-fixed backbone, separately trained residual and gate	
Training	Algorithm 1; AdamW; up to 20 epochs; patience 5
Gate	Two-layer MLP, hidden dim 16, L1 penalty
Other contextual residual settings (not the strict factorial)	
Node path	Random negatives, BCE or BPR, backbone gradients enabled
Pair path	Historical hard negatives, pair-count residual features
Stop-gradient	Stabilizer; decoupled variant close on review/wiki, lower on comment
Hard-neg ratio	$\alpha \in \{0.2, 0.5\}$ for residual sweeps
Separate adaptive probe policy	
Adaptive DTLP diagnostic policy (not PAS/PDRE)	$\alpha(\hat{\rho}_{\text{train}}) = \alpha_{\max}\max(0, 1 - \hat{\rho}_{\text{train}}/\rho_0)$ with $\alpha_{\max}$=0.2, ρ_0=0.1

E.3 Cross-backbone compatibility

Architectural prerequisite. PAS assumes separable node-general and pair-specific terminal pathways. On single-path CRAFT, sampler mixing is not pathway alignment and gives mixed outcomes: -4.5% on review, -9.9% on comment, and $+3.2\%$ on wiki. These contextual cells do not identify the internal cause. Vanilla PAS is not recommended for the nonseparable (natively entangled) CRAFT construction: contextual sampler mixing there gave mixed outcomes (Table 31), and a strict factorial on natively entangled scorers, native memory-based adaptations, and adaptations for single-path scorers that have no separate pair head remain untested.

Table 31: Contextual cross-backbone compatibility checks; only the strict TGAT factorial (Table 2) carries the primary claim. The Baseline column reports values from the cited studies, except CRAFT-wiki, which is our run (Table 38); With Plug-in reports plug-in cells. Table 32 gives matched reruns under that protocol. For DyGFormer-wiki, the published reference and plug-in are both $0.798(4)$, while the matched PDRE-protocol rerun is $0.790(4) \rightarrow 0.798(4)$. The TPNet PAS rows retain their reported historical plug-in names and per-stream constructions (Table 36), distinct from the strict factorial. Gray deltas have $|\Delta| < 2\%$. Flight/coin CRAFT-R baselines retain the source collision-check protocol; no protocol-matched delta is computed for those two rows (Appendix F.1.2).

Backbone	Dataset	Intervention	Baseline	With Plug-in	Ref. Δ
TGAT	tgbl-review	PDRE	0.355 ± 0.012	0.515 ± 0.001	$+45\%$
GraphMixer	tgbl-review	PDRE	0.521 ± 0.015	0.531 ± 0.015	$+1.9\%$
DyGFormer	tgbl-review	PDRE	0.224 ± 0.015	0.483 ± 0.015	$+116\%$
DyGFormer	tgbl-wiki	PDRE gate	0.798 ± 0.004	0.798 ± 0.004	$\approx 0\%$
TPNet	tgbl-wiki	PAS	0.827 ± 0.001	0.830 ± 0.002	$+0.4\%$
TPNet	tgbl-flight	PAS	0.884 ± 0.001	0.936 ± 0.002	$+5.9\%$
CRAFT-R	tgbl-flight	CRAFT+PAS	0.914 ± 0.000	0.923 ± 0.002	—
TPNet	tgbl-comment	PAS	0.825 ± 0.006	0.945 ± 0.015	$+14.5\%$
TPNet	tgbl-coin	PAS	0.832 ± 0.001	0.939 ± 0.002	$+12.9\%$
TPNet	Flickr	PAS	0.778 ± 0.005	0.791 ± 0.005	$+1.6\%$
CRAFT	tgbl-review	sampler mixing	0.418 ± 0.001	0.399 ± 0.012	$\mathbf{-4.5\%}$
CRAFT	tgbl-comment	sampler mixing	0.917 ± 0.006	0.826 ± 0.015	$\mathbf{-9.9\%}$
CRAFT	tgbl-wiki	sampler mixing	0.504 ± 0.005	0.520 ± 0.012	$+3.2\%$
CRAFT-R	tgbl-coin	CRAFT+PAS	0.885 ± 0.003	0.912 ± 0.002	—

Table 31 collects contextual plug-in results. Several review rows improve, GraphMixer is inside the $|\Delta| < 2\%$ band against the published reference, and the wiki rows in Table 31 are neutral or positive, whereas the matched record-order reruns in Table 32 include an adverse TGAT-wiki cell, so the wiki outcome depends on construction. TGAT–Pathway-Decoupled Residual Expert (PDRE) is a residual expert whose backbone is held fixed, distinct from PAS. On tgbl-review it scores 0.515, versus 0.4855 ± 0.0008 for the strict same-backbone PAS (RH) cell (Table 2); the two cells use different protocols (legacy record-order vs. strict timestamp-group) and are not directly comparable, and the comparison is stated only to keep PDRE and PAS distinct.

E.4 Contextual Cross-Architecture Diagnostics

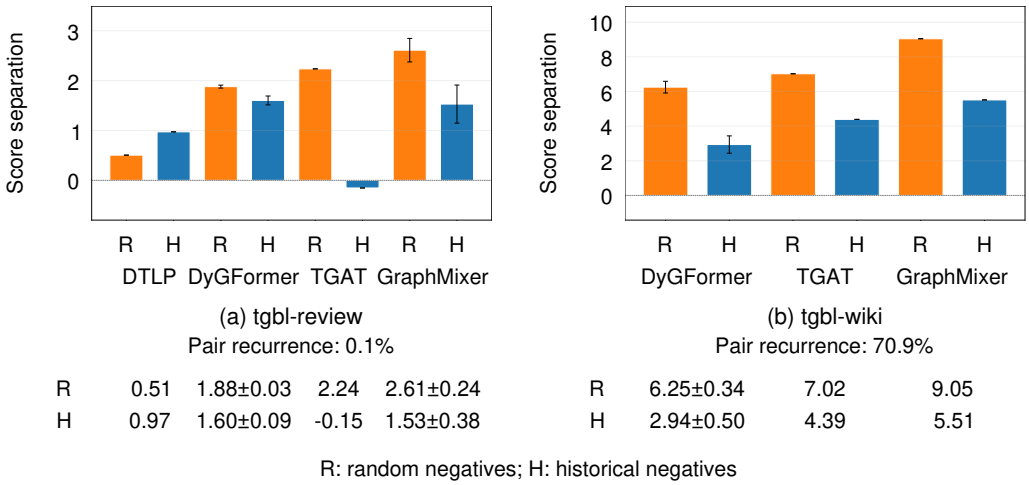
Earlier record-order PDRE experiments compare residual variants while holding each backbone fixed. They provide contextual evidence that a separately trained pair residual can interact differently with several backbones, but they are not the strict-history factorial of Table 2 and do not identify a cross-architecture causal effect.

Table 32 reports within-column comparisons from those matched record-order Pathway-Decoupled Residual Expert (PDRE) experiments, which hold the backbone fixed and use a residual pair head. PDRE is distinct from PAS and is not a co-trained sampler-to-pathway assignment. The backbone-only rows, including the DTLP rerun baseline, are reruns under that same protocol, distinct from the primary strict sampler-assignment factorial (Table 2) and the retrospective recurrence and sufficient-event audits (Tables 52 and 1). Official leaderboard baselines appear separately in Table 38. The TGAT-review PDRE cell is the learned-gate setting in Table 34; TGAT-wiki is a separate stress test.

Table 32 reports MRR for the matched record-order residual-expert comparisons. Figure 18 answers a different diagnostic question: it compares positive–negative score separation under random and historical-negative training within each displayed backbone and dataset. Its bars are not the PDRE MRR cells in the table, and absolute score gaps are not calibrated across backbones. Both objects retain their local comparison targets rather than jointly identifying the strict assignment effect.

Table 32: Matched record-order PDRE diagnostics, compared within each dataset column. These are contextual results, not strict-history causal estimates.

Method	**tgbl-wiki** (ρ=70.9%)	**tgbl-review** (ρ=0.1%)
DTLP rerun baseline	0.721 ± 0.003	0.214 ± 0.005
DTLP rerun + hard-neg	0.624 ± 0.005	0.497 ± 0.001
TGAT backbone-only rerun	0.546 ± 0.005	0.355 ± 0.012
TGAT + hard-neg (PDRE protocol)	0.495 ± 0.005	0.132 ± 0.005
TGAT + PDRE	0.508 ± 0.005	0.515 ± 0.001
GraphMixer backbone-only rerun	0.576 ± 0.005	0.394 ± 0.014
GraphMixer + hard-neg	0.563 ± 0.005	0.506 ± 0.014
GraphMixer + PDRE	0.579 ± 0.005	0.531 ± 0.015
DyGFormer backbone-only rerun	0.790 ± 0.004	0.224 ± 0.015
DyGFormer + PDRE	0.798 ± 0.004	0.483 ± 0.015



R	0.51	1.88±0.03	2.24	2.61±0.24
H	0.97	1.60±0.09	-0.15	1.53±0.38

R	6.25±0.34	7.02	9.05
H	2.94±0.50	4.39	5.51

R: random negatives; H: historical negatives

Figure 18: Positive–negative score separation ($\mathrm{pos_mean} - \mathrm{neg_mean}$) in earlier record-order diagnostic runs. Bars compare random and historical-negative training within each displayed backbone and dataset; they are not PDRE MRR cells. Absolute score gaps across backbones have no common calibrated scale. Uncertainty marks retain the reported variation; their repetition unit is unspecified. The strict factorial in Table 2 provides the causal sampler-to-pathway evidence.

E.5 Retrofitting a separable pair pathway onto a single-path backbone

Table 31 reports adverse contextual sampler-mixing outcomes for CRAFT on the unseen-dominated low-ρ review/comment data; the primary matched sampler-assignment comparison is the strict TGAT factorial in Table 2. CRAFT is a single cross-attention scorer with no isolated pair head, so native branch-wise PAS assignment is undefined. Here we test a separate external-pathway construction. Rather than reroute negatives through CRAFT's existing computation, we attach a pair pathway from the outside and leave the backbone untouched.

Concretely, we hold a trained CRAFT model fixed and add an external residual scorer $r_{\text{pair}}(\mathbf{x}_p)$ together with a learned scalar gate $g(s, d, t)$, scoring

$$\text{score}(s, d, t) \;=\; \underbrace{f_{\text{CRAFT}}(s, d, t)}_{\text{held fixed}} \;+\; g(s, d, t) \;\cdot\; r_{\text{pair}}\big(\mathbf{x}_p(s, d, t)\big), \tag{17}$$

where r_{pair} (5,313 parameters) is trained under BPR with historical hard negatives while the backbone scores enter only as a detached constant, and the gate (81 parameters) is fit on held-out queries. The backbone is never updated. Equation (17) adds a gated residual to a fixed backbone. Proposition 1 applies to an ungated residual-training objective under its stated assumptions; identical residual inputs alone do not establish cancellation for the gated score when the positive and negative gates differ. The retrofit still isolates the external residual from backbone optimization because the backbone scores are detached and the backbone is not updated.

Table 33: CRAFT residual retrofit with the backbone held fixed (test MRR; residual/gate seed 42; CRAFT backbone held fixed at seed 0). The held-fixed seed-0 model scores here, 0.7042 on wiki and 0.2583 on review, and the corresponding CRAFT baseline cells in Table 31, $0.504(5)$ and $0.418(1)$, respectively, come from separate runs and are not directly comparable. *+adapter* adds Eq. (17); the capacity/training-matched *Placebo* uses pooled-shuffled, label-decorrelated pair features. $\bar{g}$: mean test-time gate.

Dataset	ρ	Backbone	+adapter	Placebo
tgbl-wiki (seed 42)	70.9%	0.7042	0.7430 (+5.5%, $\bar{g}$=0.286)	0.7042 (+0.0%, $\bar{g}$=0.0055)
tgbl-review (seed 42)	0.1%	0.2583	0.3101 (+20.1%)	—

—: no review placebo MRR was evaluated.

The retrofit scores above the held-fixed single-path backbone in the tested cells. Table 33 reports the held-fixed-backbone scores against the same backbone plus the external pair pathway. On high-ρ tgbl-wiki the retrofit raises test MRR from 0.7042 to 0.7430 ($+5.5\%$) at seed 42 and to 0.743 ± 0.003 across five seeds (seeds 42, 123, 456, 789, and 1024, per-seed $+5.0\%$ to $+6.0\%$, mean $+5.6\%$), with the gate settling at mean $\bar{g} = 0.286$. On low-ρ tgbl-review, model-wide CRAFT sampler mixing reduces test MRR in the separately reported contextual run (-4.5%, Table 31). In the held-fixed-backbone experiment, the external retrofit improves MRR from 0.2583 to 0.3101 ($+20.1\%$) at seed 42. Across five seeds (seeds 42, 123, 456, 789, and 1024, per-seed adapter MRR $0.3101/0.3255/0.3247/0.3136/0.3206$) the retrofit reaches 0.319 ± 0.007 on tgbl-review, a mean $+23.5\%$ over the held-fixed backbone 0.2583. The contrast shows that a held-fixed-backbone pair residual can add complementary signal under this retrofit. Because the interventions differ, this comparison does not identify why model-wide CRAFT sampler mixing lowered MRR in the separately reported run.

A capacity-matched score-level placebo controls adapter capacity on wiki. The gain in Eq. (17) could come from the extra $5{,}313$ parameters rather than from any pair-level information. On wiki, the placebo column tests this alternative at score level. We train an identical residual scorer with an identical gate, but permute the pair-feature rows across the pooled set of positives and negatives so that the feature marginals are preserved while their correlation with the (s, d) label is destroyed. On tgbl-wiki this control reaches 0.7042, exactly the backbone score ($+0.0\%$), and its gate collapses to mean $\bar{g} = 0.0055$ at test time: with no usable pair signal the gate learns to switch the pathway off. The residual BPR objective is itself diagnostic. The real adapter drives residual BPR to 0.2195 on wiki and 0.1068 on review, whereas the placebo plateaus at 0.4111 on wiki and 0.6331 on review. These values separate the two training conditions but are not a calibrated no-information threshold. The score-level capacity placebo therefore supports the wiki result only; no review placebo test MRR is reported. This control varies the label-correlation of pair features at fixed capacity and training recipe, but is not a second replication of the strict TGAT factorial. A separate uniform-random residual diagnostic with different supervision is reported in Appendix E.5.1.

E.5.1 Uniform-random residual diagnostic for held-fixed CRAFT retrofit

A separate single-seed arm trains the residual with 10 uniform-random negatives per positive, rather than historical hard negatives, and reaches 0.7697 on tgbl-wiki ($+9.3\%$). Because this arm sets the

number of historical hard negatives to zero, it is neither an unfiltered-hard-negative nor a collision-filter ablation and supports no conclusion about filtering. It shows only that this seed-42 residual/gate run scored above the held-fixed CRAFT seed-0 backbone-only baseline; the individual roles of detachment, residual selection, and gating remain unresolved.

E.6 Fusion Ablation

To test the learned residual gate and the integration of the pathways, we run a fusion ablation under the Pathway-Decoupled Residual Expert (PDRE) evaluation protocol. As presented in Table 34, we evaluate various gating configurations against the backbone-only baselines. MMNS is the same pathway decoupling without the learned residual gate.

Table 34: Fusion ablation analysis under the PDRE evaluation protocol. MMNS denotes Mechanism-Matched Negative Sampling without the learned residual gate, and PDRE variants are evaluated across different gating configurations.

Method	**TGAT review**	**DyGFormer wiki**
Backbone only	0.355 ± 0.012	0.804 ± 0.005
Linear fusion	0.208 ± 0.006	0.592 ± 0.007
MMNS (no gate)	0.491 ± 0.005	0.807 ± 0.003
PDRE g=0.1	0.465 ± 0.006	0.799 ± 0.004
PDRE g=0.5	0.448 ± 0.004	0.762 ± 0.006
PDRE g=1.0	0.426 ± 0.007	0.698 ± 0.005
PDRE learned gate	0.515 ± 0.001	0.798 ± 0.004

The empirical results show that a naive linear fusion of pathways degrades performance, with a 41% decrease on the TGAT review task and a 26% decrease on the DyGFormer wiki task relative to the respective backbone-only baselines. On TGAT-review, the learned residual gate gives the best fusion result. On DyGFormer-wiki, the no-gate MMNS variant is slightly higher, so the learned gate should be read as a conservative stabilizer rather than a universally optimal combiner. The DyGFormer backbone-only baseline (0.804) is the matched rerun within this specific fusion protocol, distinct from both the standalone DyGFormer rerun in Table 32 (0.790) and the standard temporal link prediction reference value in Table 38.

E.6.1 PDRE gate placebos on tgbl-wiki (DyGFormer)

Four placebo variants against the same DyGFormer backbone (test MRR 0.804) check whether the learned PDRE gate uses the residual-expert signal rather than backbone-score recalibration alone:

- Placebo A (shuffled residual within batch): **0.7572** (-5.8% vs. backbone).
- Placebo B (random-negative-trained residual): 0.8051 ($+0.1\%$).
- Placebo C (count-only gate, no learned scores): 0.7638 (-5.0%).
- Placebo D (calibrator-only gate, $\sigma(\mathrm{MLP}(s_{\mathrm{backbone}}))$): 0.7346 ($-8.6\%$).

Three of four placebos hurt the backbone, while the random-negative residual is near the backbone score. These single-seed diagnostics show that gate calibration alone is insufficient under this evaluation protocol; they do not isolate a universal active ingredient.

E.7 Ungated Pathway-Decoupling Check

Table 35 retains the same-grid backbone and MMNS comparisons, separate from the learned-gate fusion grid.

Table 35: Mechanism-Matched Negative Sampling (MMNS) pathway-decoupling ablation. MMNS follows the decoupling direction on low-ρ review/comment but lacks the residual gate used in PDRE.

Backbone	**Method**	**wiki**	**review**	**comment**
DyGFormer	backbone only	0.790 ± 0.004	0.224 ± 0.015	0.670 ± 0.001
DyGFormer	+ MMNS pair head	0.807 ± 0.001	0.445 ± 0.009	0.812 ± 0.006
TGAT	backbone only	0.546 ± 0.005	0.355 ± 0.012	0.562 ± 0.006
TGAT	+ MMNS pair head	0.430 ± 0.006	0.491 ± 0.005	0.804 ± 0.006
DTLP	standalone selected	0.740 ± 0.004	0.497 ± 0.001	0.903 ± 0.001

Review/comment improve for both shown backbones, but wiki differs: the TGAT MMNS cell declines while the DyGFormer cell rises. This construction dependence is retained alongside the distinct learned-gate results, rather than summarized as a universal decoupling gain.

E.8 TPNet-Based Prediction and Pair-Head Variants

Table 36: TPNet-based TGB-native test MRR (reported mean $\pm$ SD), using per-stream constructions distinct from the strict factorial. References retain their cited models and protocols. The review row compares a TPNet-based construction with GraphMixer, not a modification of GraphMixer. Configuration mappings for comment, coin and flight remain unresolved in the retained records; these rows provide prediction context. Appendix F gives candidate-matched CRAFT comparisons for comment and flight.

Dataset	Reference model	Reference MRR	TPNet-based construction
tgbl-wiki	TPNet	0.827 ± 0.001	0.830 ± 0.002
tgbl-review	GraphMixer	0.521 ± 0.015	0.539 ± 0.003
tgbl-comment	TPNet	0.825 ± 0.006	0.945 ± 0.015
tgbl-flight	TPNet	0.884 ± 0.001	0.936 ± 0.002
tgbl-coin	TPNet	0.832 ± 0.001	0.939 ± 0.002

The TPNet-based prediction table retains the original reported values. The recovered review/wiki checks identify their selected configurations; the exact run and configuration mappings for the headline comment, coin and flight summaries have not been recovered. These rows provide reported predictive-performance context, rather than a verified transplant of the present strict PAS construction.

Proposition 1 applies exactly to count-residual pair pathways. This legacy single-seed TPNet variant study instead uses a richer head $h_{\text{pair}}(\mathbf{z}_s, \mathbf{z}_d, \mathbf{x}_p)$ with detached backbone embeddings and different selected configurations on the two anchors (trained with joint-random supervision on wiki and initialized from the decoupled condition with backbone gradients stopped on review). The count-only, mixed and embedding-only heads provide contextual feature comparisons, not a strict factorial.

Table 37 gives the legacy single-seed pair-head variants on the selected review and wiki anchors.

Table 37: Legacy single-seed pair-head variants on two selected anchors. The wiki and review columns use different contextual configurations and are not the strict factorial; bold marks the best variant within each column.

Pair head variant	**tgbl-review test MRR**	**tgbl-wiki test MRR**
$r_{\text{pair}}(\mathbf{x}_p)$ count-only	0.5434	0.8166
$h_{\text{pair}}(\mathbf{z}_s, \mathbf{z}_d, \mathbf{x}_p)$ default	0.5383	0.8301
$h_{\text{pair}}(\mathbf{z}_s, \mathbf{z}_d)$ embedding-only	0.4321	0.8227

On the two selected anchors, the count-only and mixed heads are within 0.006 on review, while the mixed head is 0.014 above count-only on wiki. These observations motivate retaining both feature types in exploratory variants but do not establish a recurrence-determined feature transition, and they do not extend Proposition 1 beyond the count residual.

F BENCHMARK RESULTS BY TASK AND CANDIDATE PROTOCOL

Dataset Coverage. The full study covers 16 datasets[1] spanning Temporal Graph Benchmark (TGB) link prediction, TGB 2.0 temporal knowledge and heterogeneous graphs (TKG/THG), and the DyGLib benchmark suite (Yu et al., 2023) for broader mechanism and scope diagnostics. Appendices also report TGB-Seq (100 negatives) and TGB node-property prediction (NDCG@10) (Huang et al., 2023; Gastinger et al., 2024; Yu et al., 2023; Yi et al., 2025b;a; Lu et al., 2024).

Contextual reference methods. For the probe and compatibility tables, reference values are taken from TGB (Huang et al., 2023), TGB 2.0 (Gastinger et al., 2024), DyGLib/DyGFormer (Yu et al., 2023), TGB-Seq (Yi et al., 2025b), CRAFT (Yi et al., 2025a), TPNet (Lu et al., 2024), GraphMixer (Cong et al., 2023), TGN (Rossi et al., 2020), TGAT (Xu et al., 2020), EdgeBank (Poursafaei et al., 2022), TLogic (Liu et al., 2022), NAVIS (Feldman et al., 2026), and heuristic sources.

Abbreviations are BPR (Bayesian Personalized Ranking), BCE (binary cross-entropy), CE (cross-entropy), TKG (temporal knowledge graph), and THG (temporal heterogeneous graph).

The following families retain their own candidate construction, metric and model-selection rules. Each grid is interpreted separately; the cross-protocol summary supplies context, not pooled ranks.

F.1 TGB LINK PREDICTION

The diagnostic rows have different responses even within TGB-native evaluation: historical negatives raise the probe scores on review/comment but lower wiki, while the additional high-recurrence rows include positive, near-null and adverse effects. Published baselines, fixed-policy probes and validation-selected probes retain their captioned configurations. The following fixed-policy and matched-protocol checks therefore remain separate comparisons.

F.1.1 TGB-NATIVE LINK-PREDICTION COMPARISONS

Table 40 reports the fixed and routed TGB-native link-prediction comparison rows under their captioned configurations.

The tgbl-comment TPNet+PAS gain is 3.0% against the global CRAFT prior; the larger Table 31 delta is same-backbone. The tgbl-wiki PRF and hard-negative-sampling rows apply distinct interventions and do not identify separate failure modes.

F.1.2 PROTOCOL-MATCHED COMPARISON ON tgbl-flight

The paper's tgbl-flight cell for TPNet+PAS uses TGB-native evaluation (no collision check), while CRAFT-R (0.9139) uses 100 random negatives plus collision checking. A single-seed rerun (seed 42) under the CRAFT-R protocol gives 0.9195, $+0.6$ points over that baseline. This matched rerun, rather than the higher native-protocol cell, is the only direct flight comparison.

F.1.3 DTLP UNDER THE CRAFT-R EVALUATION PROTOCOL

We re-evaluate the seed-42 DTLP model on tgbl-review under CRAFT-R's collision-checked random-negative protocol, training on the same data with the same hyperparameters and selecting the model with the best validation MRR. Table 41 shows a -5.1% change in this run. For tgbl-coin, $0.912(2)$ is CRAFT-R+PAS, whereas 0.939 is TPNet+PAS under TGB-native evaluation; no same-model protocol contrast is available.

[1]The count comprises the five main TGB-LP streams (tgbl-wiki, tgbl-review, tgbl-comment, tgbl-flight, and tgbl-coin), the five non-legacy TGB 2.0 streams in Table 42 (thgl-github, thgl-myket, thgl-software, tkgl-smallpedia, and tkgl-polecat), and the six DyGLib streams in Table 44. The legacy tkgl-yago stress-test column and appendix-only rows are excluded.

Table 38: MRR on TGB link-prediction benchmarks. Upper block: published baselines except CRAFT-wiki[‡], which is a run performed for this paper; starred flight/coin CRAFT-R cells retain the source collision-check protocol and are context only; starred cells in the CRAFT variants row use the CRAFT-R backbone/protocol. Lower block: our DTLP runs and reported TPNet+PAS plug-in variants, distinct from the strict factorial, rather than a common leaderboard. DTLP uses streaming statistics (Appendix D.7 gives the fixed-statistic control). In the CRAFT variants row, the wiki, review, and comment cells report model-wide sampler mixing in the single-path CRAFT scorer; these cells do not instantiate branch-wise PAS. An em dash denotes a value not reported by the cited source or outside our evaluated grid.

Method	**tgbl-wiki**	**tgbl-review**	**tgbl-comment**	**tgbl-flight**	**tgbl-coin**
CRAFT	$(0.504 \pm 0.005)^{\ddagger}$	0.418 ± 0.001	0.917 ± 0.006	$(0.914 \pm 0.000)^{\star}$	$(0.885 \pm 0.003)^{\star}$
TPNet	0.827 ± 0.001	—	0.825 ± 0.006	0.884 ± 0.001	0.832 ± 0.001
Heuristic	0.821	—	0.723	0.878	0.774
HyperEvent	0.810 ± 0.002	—	0.760 ± 0.002	0.877 ± 0.003	0.773 ± 0.002
DyGFormer	0.798 ± 0.004	0.224 ± 0.015	0.670 ± 0.001	—	0.752 ± 0.004
NAT	0.749 ± 0.010	0.341 ± 0.020	—	—	—
Base3	0.743	0.108	0.449	—	—
DyGMamba	0.739 ± 0.009	—	—	—	—
TNCN	0.718 ± 0.001	0.377 ± 0.010	0.697 ± 0.006	0.820 ± 0.004	0.762 ± 0.004
CAWN	0.711 ± 0.006	0.193 ± 0.001	—	—	—
CTAN	0.668 ± 0.007	0.405 ± 0.004	0.671 ± 0.067	—	0.748 ± 0.004
EdgeBank(tw)	0.571	0.025	0.149	0.387	0.580
EdgeBank(unl.)	0.495	0.023	0.129	0.167	0.359
HTGN(UTG)	0.464 ± 0.005	0.104 ± 0.002	—	—	—
TGN	0.396 ± 0.060	0.349 ± 0.020	0.379 ± 0.021	0.705 ± 0.020	0.586 ± 0.037
EGCNo(UTG)	0.398 ± 0.007	0.195 ± 0.001	—	—	—
GCLSTM(UTG)	0.374 ± 0.010	0.095 ± 0.002	—	—	—
GCN(UTG)	0.336 ± 0.009	0.186 ± 0.002	—	—	—
TCL	0.207 ± 0.025	0.193 ± 0.009	—	—	—
TGAT	0.141 ± 0.007	0.355 ± 0.012	—	—	—
GraphMixer	0.118 ± 0.002	0.521 ± 0.015	—	—	—
DyRep	0.050 ± 0.017	0.220 ± 0.030	0.289 ± 0.033	0.556 ± 0.014	0.452 ± 0.046
DTLP valid.-selected (no PRF)	0.740 ± 0.004	0.497 ± 0.001	0.903 ± 0.001	0.884 ± 0.003	0.773 ± 0.005
DTLP fixed + PRF	0.744 ± 0.005	0.501 ± 0.003	0.919 ± 0.001	0.885 ± 0.003	0.829 ± 0.002
DTLP full (random)	0.721 ± 0.003	0.223 ± 0.009	0.675 ± 0.010	0.883 ± 0.001	0.773 ± 0.005
DTLP full + hist. hard-neg	0.624 ± 0.005	0.497 ± 0.001	0.903 ± 0.001	0.893 ± 0.005	0.808 ± 0.005
DTLP pair-only	0.698 ± 0.001	0.175 ± 0.003	0.392 ± 0.005	0.879 ± 0.005	0.724 ± 0.006
DTLP entity-only	0.060 ± 0.001	0.361 ± 0.001	0.744 ± 0.002	0.631 ± 0.005	0.725 ± 0.004
TPNet + PAS	0.830 ± 0.002	0.539 ± 0.003	0.945 ± 0.015	0.936 ± 0.002	0.939 ± 0.002
CRAFT variants	0.520 ± 0.012	0.399 ± 0.012	0.826 ± 0.015	$(0.923 \pm 0.002)^{\star}$	$(0.912 \pm 0.002)^{\star}$

Table 39: Additional TGB-native datasets not used in the primary three-dataset same-backbone runs; rows use the native no-collision-check protocol unless noted.

Dataset	ρ **(%)**	**MRR**	**Notes**
tgbl-coin	85.8	0.773 ± 0.005	DTLP validation-selected / no PRF; +PRF reaches 0.829 ± 0.002. High ρ but hard-neg helps ($+4.5\%$); 638K nodes
thgl-forum	26.9	0.804 ± 0.005	TGN(edge-type) reference: 0.729
tgbl-subreddit	86.5	0.737 ± 0.001	High ρ, hard-neg neutral (-0.6%)
tgbl-enron	97.8	0.516 ± 0.004	Very high ρ, hard-neg mildly hurts (-3.1%)
tgbl-uci	65.7	0.326 ± 0.003	Mid ρ, hard-neg strongly hurts (-34.0%)
tgbl-lastfm	87.1	0.120 ± 0.001	High ρ, hard-neg neutral ($+0.8\%$)

Table 40: TGB-native link-prediction comparison rows. DTLP fixed uses a dataset-specific TGB-native policy; the DTLP validation-selected entry in Table 38 is selected separately over loss/lr and sampler settings. Fixed and reported plug-in entries use no collision check; starred CRAFT/CRAFT-R entries retain their source protocol as best-prior context. The TPNet+PAS name denotes the reported historical plug-in variants, distinct from the strict factorial.

Dataset	Role	Method / construction	Test MRR
tgbl-comment	Fixed probe	DTLP	0.903 ± 0.001
	Recency feature	DTLP+ PRF	0.919 ± 0.001
	Reference	TPNet	0.825 ± 0.006
	Reported plug-in	TPNet+PAS	0.945 ± 0.015
	Best prior	CRAFT	0.917 ± 0.006
tgbl-wiki	Fixed probe	DTLP	0.721 ± 0.003
	Recency feature	DTLP+ PRF	0.744 ± 0.005
	Reference	TPNet	0.827 ± 0.001
	Reported plug-in	TPNet+PAS	0.830 ± 0.002
	Best prior	TPNet	0.827 ± 0.001
tgbl-review	Fixed probe	DTLP	0.497 ± 0.001
	Recency feature	DTLP+ PRF	0.501 ± 0.003
	Reference	GraphMixer	0.521 ± 0.015
	Reported plug-in	TPNet+PAS	0.539 ± 0.003
	Best prior	GraphMixer	0.521 ± 0.015
tgbl-coin	Fixed probe	DTLP	0.808 ± 0.005
	Recency feature	DTLP+ PRF	0.829 ± 0.002
	Reference	CRAFT-R	$0.885^{\star} \pm 0.003$
	Reported plug-in	TPNet+PAS	0.939 ± 0.002
	Best prior	CRAFT-R	$0.885^{\star} \pm 0.003$
tgbl-flight	Fixed probe	DTLP	0.884 ± 0.003
	Recency feature	DTLP+ PRF	0.885 ± 0.003
	Reference	TPNet	0.884 ± 0.001
	Reported plug-in	TPNet+PAS	0.936 ± 0.002
	Best prior	CRAFT-R	$0.914^{\star} \pm 0.000$

Table 41: DTLP seed 42 on tgbl-review: TGB-native protocol vs. the CRAFT-R protocol with collision-checked random negatives and a model selected by validation MRR.

Dataset	TGB-native MRR	CRAFT-R MRR	Δ
tgbl-review	0.497	0.4716	-5.1%

F.2 TGB 2.0 Knowledge and Heterogeneous Graphs

The dictionary-based probe is above EdgeBank on github but below STHN on software; the relative standing therefore varies across the displayed graph types. The legacy yago comparator and dictionary-based yago score use different query/candidate definitions and are not a matched gap. The extra icews, wikidata and forum rows extend coverage under their stated dictionaries.

For the representative yago stress test, the legacy 40K-query pair-only margin-loss probe scores 0.875 ± 0.018 against our matched Repeat/Recurrency run (0.803); the dictionary-based probe instead scores 0.916 ± 0.015. CRAFT-R denotes the recurrence variant of CRAFT (Yi et al., 2025a), and STHN is a temporal heterogeneous-network model (Li et al., 2023). These reference names do not remove the row-specific protocol distinctions.

Table 42: MRR on TGB 2.0 TKG/THG benchmarks plus a legacy tkgl-yago stress-test column. EdgeBank, TGN/TGN(edge-type), CEN, TLogic, RE-GCN, and STHN values are taken from the TGB 2.0 source where reported (Gastinger et al., 2024); ±std is shown when available. Cells marked "OOM" exceeded available GPU memory. Recurrency/Repeat rows marked ‡ are our own heuristic runs, not published leaderboard entries. The DTLP and TPNet+PAS rows are also our runs under the official precomputed negative-candidate dictionaries. These are contextual model comparisons, not estimates of a model-class effect; a gap is interpretable only when its entries share the same evaluation protocol. In the tkgl-yago column, Repeat/Recurrency is matched to the legacy 40K-query DTLP result in Table 19, not to the dictionary-based DTLP cell shown here. An em dash denotes a value not reported by the cited source or outside our evaluated grid.

Method	**thgl-github**	**thgl-myket**	**thgl-software**	**tkgl-yago**	**tkgl-smallpedia**	**tkgl-polecat**
EdgeBank (unlimited)	0.413	0.456	0.449	0.478	0.333	0.045
EdgeBank(tw)	0.374	0.245	0.288	0.617	0.353	0.056
Recurrency Baseline (train)‡	—	—	—	—	0.605	0.198
Recurrency Baseline (default)‡	—	—	0.099	0.858	0.486	0.167
Repeat/ Recurrency‡	0.101	0.427	0.093	0.803	0.321	—
Popularity‡	0.291	0.893	—	0.021	0.030	—
Combined‡	0.341	0.901	0.158	0.737	0.269	—
TGN	OOM	OOM	0.324 ± 0.017	—	—	—
TGN(edge-type)	OOM	OOM	0.424 ± 0.013	—	—	—
CEN	OOM	OOM	—	—	0.612 ± 0.001	0.184 ± 0.002
TLogic	—	—	—	—	0.595 ± 0.001	0.228 ± 0.001
RE-GCN	—	—	—	—	0.594 ± 0.001	0.175 ± 0.002
STHN	OOM	OOM	0.731 ± 0.005	—	—	—
TPNet + PAS	—	0.938 ± 0.005	0.832 ± 0.005	0.611 ± 0.001	0.603 ± 0.048	0.645 ± 0.001
DTLP	0.818 ± 0.007	0.904 ± 0.002	0.547 ± 0.003	0.916 ± 0.015	0.774 ± 0.029	0.229 ± 0.016

Table 43: Additional TGB 2.0 TKG/THG coverage under the stated precomputed negative-candidate dictionaries. DTLP is a streaming diagnostic probe, not a common leaderboard entry or evidence that a small count model solves general TLP. Best baseline numbers are from the TGB 2.0 leaderboard (Gastinger et al., 2024); Δ is a within-protocol gap only. Displayed ρ describes data; it does not select samplers.

Dataset	ρ **(%)**	**DTLP**	**Best baseline**	Best baseline's MRR	Δ
tkgl-icews	51.9	0.395 ± 0.008	Recurrency Baseline (train)	0.211	+87%
tkgl-wikidata	13.0	0.864 ± 0.005	EdgeBank(tw)	0.535	+61%
thgl-forum	26.9	0.804 ± 0.005	TGN(edge-type)	0.729 ± 0.009	+10%

F.3 DyGLib Candidate Protocols

DyGFormer/CRAFT-R/CAWN/SGNN-HN/TGN/JODIE/GraphMixer/TGAT/DyRep are from Yi et al. (2025a) Table 3 under 100 random negatives plus collision checking. The entire TPNet‡ row consists of our own matched-protocol runs; TPNet is not evaluated in that source. Our DTLP row uses the same candidate and collision-check protocol. An em dash denotes a value not reported by the cited source; in the TPNet row it denotes lastfm, which did not converge under this protocol. †mooc has only 97 unique destinations, so we evaluate with 20 negative candidates; this protocol difference is disclosed and the cell is not used for a cross-method claim. SocialEvo is excluded because its small pool forces fallback under fixed-k sampling.

Table 44: MRR on DyGLib benchmarks. Comparison sources, candidate construction, and the row-specific missing-entry and small-pool exceptions are specified in the text above.

Method	wikipedia	reddit	mooc[†]	lastfm	uci	Flights
DyGFormer	0.887 ± 0.000	0.887 ± 0.001	0.422 ± 0.013	0.465 ± 0.005	0.766 ± 0.003	0.841 ± 0.035
CRAFT-R	0.883 ± 0.003	0.893 ± 0.001	0.623 ± 0.004	0.552 ± 0.002	0.751 ± 0.001	0.832 ± 0.002
CAWN	0.861 ± 0.001	0.877 ± 0.000	0.270 ± 0.006	0.351 ± 0.003	0.670 ± 0.005	0.801 ± 0.021
SGNN-HN	0.822 ± 0.004	0.874 ± 0.000	0.572 ± 0.002	0.407 ± 0.003	0.572 ± 0.004	0.808 ± 0.001
TGN	0.822 ± 0.004	0.793 ± 0.002	0.390 ± 0.045	0.256 ± 0.044	0.453 ± 0.064	0.642 ± 0.021
JODIE	0.765 ± 0.017	0.772 ± 0.013	0.199 ± 0.031	0.185 ± 0.029	0.514 ± 0.017	0.204 ± 0.042
GraphMixer	0.736 ± 0.015	0.709 ± 0.004	0.290 ± 0.003	0.268 ± 0.014	0.596 ± 0.010	0.423 ± 0.005
TGAT	0.727 ± 0.015	0.780 ± 0.003	0.316 ± 0.041	0.241 ± 0.013	0.315 ± 0.005	0.539 ± 0.011
DyRep	0.674 ± 0.042	0.723 ± 0.021	0.177 ± 0.013	0.178 ± 0.066	0.140 ± 0.027	0.156 ± 0.005
TPNet[‡]	0.906 ± 0.003	0.823 ± 0.008	0.695 ± 0.009	—	0.808 ± 0.004	0.891 ± 0.006
DTLP	0.901 ± 0.004	0.945 ± 0.002	0.596 ± 0.022	0.912 ± 0.002	0.811 ± 0.004	0.895 ± 0.002

Table 45: DTLP loss variants on DyGLib benchmarks under the standard DyGLib evaluation protocol (full-destination MRR with no collision check). Tables 44 and this table report separately recorded protocols, without establishing a fixed-checkpoint, matched-query comparison. Their cross-table MRR differences are descriptive and cannot be attributed solely to candidate construction.

Dataset	ρ (%)	$\lvert V \rvert$	BPR	CE	Margin
mooc	59.8	7.1K	0.911 ± 0.006	0.910 ± 0.005	0.912 ± 0.006
Flights	86.7	13.2K	0.847 ± 0.004	0.849 ± 0.005	0.847 ± 0.004
wikipedia	89.8	9.2K	0.948 ± 0.003	0.946 ± 0.004	0.947 ± 0.003
reddit	91.5	11.0K	0.975 ± 0.001	0.975 ± 0.002	0.975 ± 0.001
lastfm	87.1	1.9K	0.597 ± 0.003	0.601 ± 0.004	0.600 ± 0.004
uci	65.7	1.9K	0.732 ± 0.003	0.748 ± 0.004	0.750 ± 0.004

The collision-checked 100-negative table and full-destination loss grid answer different questions. In the latter, CE and margin change uci more than reddit; on lastfm, the large differences across candidate protocols exceed the within-grid loss differences. The mooc candidate-count exception remains excluded from cross-method ranking.

F.4 TGB-SEQ

Within the 100-negative probe grid, the full BPR/CE probe exceeds entity-only on ML-20M and Taobao, whereas the GoogleLocal CE and entity-only means coincide. The published-method subset is likewise mixed: the probe is close to CRAFT on Flickr but lower on WikiLink. These task-specific comparisons do not turn the streaming probe into a general model-class claim.

Table 46: MRR for DTLP on TGB-Seq benchmarks under the 100-negative evaluation protocol. The entity-only comparison describes sensitivity to component inclusion under this protocol; it does not establish a mechanism for the ranking differences.

Dataset	BPR	CE	Entity-only
Flickr	0.613 ± 0.004	0.616 ± 0.006	0.590 ± 0.007
GoogleLocal	0.190 ± 0.003	0.195 ± 0.004	0.195 ± 0.004
ML-20M	0.262 ± 0.004	0.269 ± 0.006	0.180 ± 0.003
Patent	0.163 ± 0.002	0.161 ± 0.003	0.150 ± 0.002
Taobao	0.431 ± 0.003	0.432 ± 0.004	0.305 ± 0.006
WikiLink	0.571 ± 0.006	0.586 ± 0.005	0.504 ± 0.004
Yelp	0.378 ± 0.005	0.380 ± 0.006	0.320 ± 0.004
YouTube	0.581 ± 0.004	0.612 ± 0.006	0.580 ± 0.005

Table 47: TGB-Seq published-method comparison under the 100-negative evaluation protocol. Baseline values (CRAFT, DyGFormer, GraphMixer) and their standard deviations are taken directly from (Yi et al., 2025a) Table 2. An em dash denotes a value outside the cited study's evaluation grid. DTLP values are our multi-seed comparison subset; the full TGB-Seq task grid is reported separately in Table 46. DTLP remains a streaming diagnostic, and the parameter counts do not establish a model-class efficiency claim.

Method	**Params**	**YouTube**	**Flickr**	**WikiLink**
CRAFT	181M	0.589 ± 0.002	0.623 ± 0.008	0.755 ± 0.008
DyGFormer	$\sim$500K	0.461 ± 0.034	0.496 ± 0.029	—
GraphMixer	$\sim$500K	0.589 ± 0.001	0.450 ± 0.001	0.486 ± 0.000
DTLP	11K	0.612 ± 0.006	0.616 ± 0.006	0.586 ± 0.005

F.5 Node-Property Prediction

Table 48: TGB node-property prediction benchmarks, NDCG@10. Baseline values are from the TGB node-property leaderboard (Huang et al., 2023); an em dash denotes an entry outside the leaderboard's evaluation grid. NAVIS, Moving Average, Persistent Forecast, and TGNv2 are reported as point values without standard deviations on the leaderboard. DTLP-node values and reported spreads are retained from the original study; their run-to-table mapping is discussed in Appendix F.5.

Method	**tgbn-reddit**	**tgbn-genre**	**tgbn-trade**	**tgbn-token**
NAVIS	0.569	0.528	0.863	0.513
Moving Average	0.559	0.509	0.823	0.508
Persistent Forecast	0.369	0.357	0.855	0.430
TGNv2	0.507	0.469	0.735	0.294
DyGFormer	0.316 ± 0.001	0.365 ± 0.002	0.388 ± 0.006	—
TGN	0.315 ± 0.020	0.367 ± 0.058	0.374 ± 0.001	0.169 ± 0.006
DyRep	0.312 ± 0.001	0.351 ± 0.001	0.374 ± 0.001	0.141 ± 0.006
DTLP-node	0.575 ± 0.003	0.529 ± 0.002	0.795 ± 0.003	0.389 ± 0.002

The node-level head is competitive on reddit/genre but below the listed strong heuristics on trade/token. The means and reported spreads are retained from the submitted table. The recovered result directory contains multiple configurations and repeated runs; the exact run-to-table selection and aggregation mapping has not been recovered, so this revision does not claim to have independently reproduced those summaries. These NDCG@10 observations concern node-property prediction; no routed node-head intervention was evaluated, and they are not pooled with link-prediction MRR.

F.6 Auxiliary Cross-Protocol Results

Table 49 collects auxiliary TLP, THG, TGB-Seq, and node-prediction rows with different protocols.

These TGB-Seq rows provide exploratory cross-protocol context; their different protocols preclude a direct ranking of the displayed methods.

G Retrospective Recurrence Context

Recurrence is an observed descriptor, not an input required by PAS. We group the dataset taxonomy, heterogeneous intervention audit, fitted heuristic and post-hoc recombination here. The toy sampler comparisons keep their separate constructed setting.

Table 49: Auxiliary non-identical-protocol rows. TLP/THG/TGB-Seq values use MRR, with TGB-Seq under the 100-negative protocol; node-prediction rows use NDCG@10. The DTLP-node entries report the means and standard deviations in Table 48; NAVIS entries remain source point estimates. Percentage deltas are computed from full-precision means. An em dash in the routed column means that no routed node-prediction result was evaluated.

Dataset / task	DTLP	Routed result	Reference
thgl-software (THG)	0.547 ± 0.003	CRAFT+PAS 0.585 ± 0.005	STHN 0.731 ± 0.005
GoogleLocal (TGB-Seq)	0.195 ± 0.004	CRAFT+PAS 0.621 ± 0.006	CRAFT 0.624 ± 0.005
Yelp (TGB-Seq)	0.380 ± 0.006	CRAFT+PAS 0.732 ± 0.004	CRAFT 0.727 ± 0.006 (+0.7%)
ML-20M (TGB-Seq)	0.269 ± 0.006	CRAFT+PAS 0.433 ± 0.006	CRAFT 0.359 ± 0.007 (+20.5%)
Taobao (TGB-Seq)	0.432 ± 0.004	CRAFT+PAS 0.709 ± 0.004	CRAFT 0.707 ± 0.004 (+0.2%)
tgbn-trade (NDCG)	0.795 ± 0.003	—	NAVIS 0.863
tgbn-token (NDCG)	0.389 ± 0.002	—	NAVIS 0.513

G.1 Pair Recurrence as a Descriptive Axis

For a stream of N positive queries, define

$$\rho_{\text{stream}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\exists j : \; t_j < t_i, \; (s_j, d_j) = (s_i, d_i)] \, . \tag{18}$$

This is the interaction-weighted homogeneous-stream repeat ratio used by TGB-Seq (Yi et al., 2025b), not a new statistic or an unweighted source average. Related pair-revisit and collision checks appear in CRAFT (Yi et al., 2025a). Equation (18) defines strict stream recurrence. Separately, the reported review, comment, and wiki diagnostic values measure the fraction of test queries whose ordered pair appeared in training or validation; they exclude recurrence first established within the test stream. These quantities are not interchangeable. Other contextual recurrence values retain their stated source windows and estimands. Bare ρ in a table denotes this quantity under the observation window declared in its caption; ρ_{diag} is retrospective diagnostic metadata, $\hat{\rho}_{\text{train}}$ is prefix-only metadata, and a warmup estimate is labeled explicitly. Recurrence can describe whether pair novelty is common. It implies no universal sampler crossover, and PAS consumes none of these quantities. Count-residual cancellation depends on query-conditioned negative-sampling probability mass outside $\mathcal{H}_s(t)$; future-positive overlap depends on the composition and horizon of the historical pool. Density, support, objective, and architecture alter both quantities. The earlier 10% split was selected post hoc from heterogeneous interventions and is retained only as exploratory context in Appendix G.3, with the limitations documented alongside that audit.

G.2 Feature Diagnostics and Dataset Taxonomy

Table 50 groups datasets by structural properties, pair recurrence, and relational complexity. The groups summarize observed benchmark characteristics. Table 51 reports pathway-level diagnostics for representative datasets.

Table 50: Descriptive dataset grouping by structural properties. Examples summarize observed cells, not a validated selection policy or deployment rule.

Regime	Characteristics	Example / observed diagnostic focus
Sparse interaction	Low recurrence, skewed degree	github, myket / entity features
Pair-repetitive	Moderate/high recurrence or strong pair-count diagnostics	software, yago / pair features
Multi-relational	High pair, low triple recurrence	polecat / relation-aware or structural feats.
Many-relation KG	Many relations, moderate recurrence	smallpedia / pair features
High-rec. TLP	Pair rec. > 0.5, one relation	tgbl-wiki / pair features
Low-rec. TLP	Pair rec. < 0.1, one relation (post-hoc group)	review, comment / tested routed or hard-neg cells

Table 51: Model diagnostics across datasets. Pair % is the fraction of the positive-negative score gap from the pair scorer.

Dataset	Pair (%)	Entity (%)	Active relations	Top pair feature
tkgl-yago	87	13	10/20	reverse-edge count (r=0.84)
thgl-software	70	30	14/14	pair count (r=0.83)

G.3 Retrospective Recurrence Audit

Table 52 preserves the retrospective audit but removes its accuracy score. The 10% split was chosen after observing heterogeneous interventions that differ in model, loss, sampler strength, and evaluation protocol; it is therefore descriptive evidence, not a held-out classifier. The tkgl-yago row uses the documented matched BPR-grid pair $0.494 \rightarrow 0.192$ (-61%), rather than the unreconstructable cross-configuration $0.827 \rightarrow 0.192$ comparison.

Table 52: Retrospective recurrence audit. Effects are within-row diagnostics but are not samples from one standardized intervention. ρ_{diag} denotes retrospective diagnostic metadata, not one common full-stream estimator; the review, comment, and wiki rows use test/pretest pair overlap as defined in Appendix G.1. These outcome-window quantities are not leakage-free deployment features.

Dataset	ρ_{diag}	$\lvert V\rvert$	Observed effect	Intervention family
tgbl-review	0.1%	353K	+123%	diagnostic pair supervision
tgbl-comment	6.1%	995K	+34%	diagnostic pair supervision
thgl-software	23.1%	682K	−5.7%	protocol-specific probe
tkgl-smallpedia	25.0%	47K	−2.7%	BPR probe
mooc	59.8%	7.1K	−85%	DyGLib intervention
tgbl-uci	65.7%	1.9K	−34%	TGB intervention
tgbl-wiki	70.9%	9.2K	−13%	TGB intervention
tkgl-yago	74.3%	10.6K	−61%	matched BPR grid
tgbl-coin	85.8%	638K	+4.5%	TGB intervention
tgbl-subreddit	86.5%	11K	−0.6%	TGB intervention
Flights (DyGLib)	86.7%	13.2K	−3.1%	DyGLib intervention
tgbl-lastfm	87.1%	1.9K	+0.8%	TGB intervention
wikipedia	89.8%	9.2K	−40%	DyGLib intervention
reddit	91.5%	11K	−19%	DyGLib intervention
tgbl-enron	97.8%	0.2K	−3.1%	TGB intervention

Table 53 gives the 30-bucket density-aware consistency check used in the retrospective recurrence audit.

Table 53: Bucket-level logistic regression over 30 pooled chronological buckets from tgbl-wiki and tgbl-review. Here ρ_b and d_b denote the bucket recurrence rate and bucket density, respectively. Both models achieve 100% in-sample agreement; this is a consistency check for the density-aware refinement, not an independent generalization estimate.

Model	Coefficients	Interpretation
$\sigma(\beta_0 + \beta_1\rho_b)$	$\beta_1 < 0$	Recurrence separates these bucket sets
$\sigma(\beta_0 + \beta_1\rho_b + \beta_2 d_b)$	$\beta_1 = -20.4,\ \beta_2 = 4.3$	Density refines the fitted boundary

G.4 Retrospective Diagnostics of Descriptors and Samplers

The following analyses ask two separate questions. The retrospective audit compares descriptor agreement with the signs of heterogeneous recorded interventions. The constructed toy instead compares samplers under a matched zero-overlap budget, asking whether larger pair-feature contrast is sufficient to yield better ranking. The two blocks of Table 54 retain their different units: dataset-level sign agreement and multi-seed toy MRR.

G.4.1 Retrospective Descriptor Agreement

Forward overlap c_d is the fraction of sampled historical negatives appearing in the future window. The historical audit covers 11 datasets: tgbl-review, comment, uci, wiki, coin, subreddit, lastfm, enron (all tgbl-), thgl-software, tkgl-smallpedia and tkgl-yago. LOO-fitted c_d thresholds yield $8/11$ utility-sign agreements; the fixed post-hoc $\rho < 10\%$ grouping yields $9/11$ without foldwise refitting. Labels mix interventions; the original audit flags alignment/degeneracy concerns in the three non-tgbl rows. These are historical agreements, not matched effects or held-out selector validation. Coin's $c_d = 0.286$ ranks fifth-highest, yet its effect is $+4.5\%$; both rules predict "hurt". In the eight tgbl rows, help $[0.0005, 0.509]$ and hurt $[0.083, 0.412]$ overlap: uci hurts at $c_d = 0.083$, lastfm helps at $c_d = 0.509$. Prefix/future Pearson correlation 0.903 covers nine paired records, excluding coin (no prefix record) and yago (degenerate overlap), not all eleven. Neither descriptor is a validated stand-alone policy.

G.4.2 Sampler Comparisons in the Controlled Toy

As an empirical probe, we test a sampler that maximizes the pair-feature gap $\|\Delta \mathbf{x}_p\|$ between positive and negative. Proposition 1 proves only residual-parameter cancellation under identical inputs and does not imply that maximizing this gap improves MRR. We use a controlled toy constructed to exhibit the qualitative low-ρ contrast under historical-negative evaluation: random negatives score 0.024 and historical hard negatives score 0.877. The toy is not a benchmark replica. At a matched zero-overlap budget (every sampler rejects later-positive candidates), historical hard negatives reach 0.879 while the feature-contrast sampler reaches 0.863 at low ρ (5 seeds), even though the feature-contrast sampler produced a $1.4\times$ larger $\|\Delta \mathbf{x}_p\|$ (0.796 vs. 0.550). Larger pair-feature contrast did not translate into better learning. At high ρ the two tie within noise once later-positive candidates are removed. In this toy, an explicit contrast objective adds no separable improvement over historical sampling; this is not a benchmark-level optimality claim.

Score difficulty in the controlled toy. A MixGCF-style score-difficulty sampler, which selects negatives by predicted score rather than by history or feature contrast, scores below historical hard and feature contrast at low ρ (0.748 vs. 0.879 for historical hard) and ties random at high ρ. Score difficulty is weaker in this construction; that does not rule it out in other models or candidate protocols.

Table 54: Retrospective descriptor agreement and controlled sampler diagnostics. *Top:* historical sign agreement for the fixed post-hoc ρ grouping and LOO-fitted forward-overlap threshold c_d; neither is a validated deployment rule. *Bottom:* low-ρ toy MRR at a matched zero-overlap budget (5 seeds), with mean pair-feature gap $\|\Delta \mathbf{x}_p\|$.

Descriptor: historical sign agreement (11 datasets)		
Descriptor	Sign agreement	coin (c_d=0.286)
Post-hoc $\rho < 10\%$ group	$9/11$	wrong (same as c_d)
c_d (forward overlap)	$8/11$	wrong
Sampler: low-ρ toy MRR, matched zero-overlap budget		
Sampler	MRR (5 seeds)	$\|\Delta \mathbf{x}_p\|$
random	0.024 ± 0.000	0.027
historical hard	0.879 ± 0.007	0.550
feature-contrast	0.863 ± 0.009	0.796
score-difficulty (MixGCF)	0.748 ± 0.032	0.042

The descriptor audit retains a concrete failure case: both rules predict harm on coin despite its positive recorded effect. In the controlled toy, the feature-contrast sampler produces a larger input gap but lower MRR than historical sampling at the matched zero-overlap budget. These are separate reasons not to treat a descriptor or a contrast magnitude as a stand-alone selection rule. They do not invalidate the favorable historical-sampling result observed in the toy.

G.5 Post-hoc Heuristic Calculation

The original 10% grouping was chosen after observing heterogeneous outcomes; it is not a predictive crossover or a held-out decision rule. This section retains the first-order calculation as a post-hoc illustration, not a measured parameter sweep. It uses free parameters β_+, β_-, κ, q_s, and η_s, with κ calibrated from the observed tgbl-wiki overlap rate.

Assumptions. First, around a trained operating point, the expected change in reciprocal rank is approximated by the expected change in pairwise BPR updates induced by the sampler. Second, if an α fraction of random negatives is replaced by historical negatives, we approximate the recovered pair-pathway update opportunity by $\alpha\eta_s(1-\rho_s)$, where $\eta_s \in [0,1]$ summarizes negative-novelty and candidate-set density. Third, only a fraction $q_s \in [0,1]$ of recovered updates matters for evaluation. Fourth, the future-overlap rate is approximated as $c_s \approx \kappa\rho_s$. Fifth, useful updates contribute scale $\beta_+ > 0$, while this heuristic assumes overlapping updates contribute an opposing scale $\beta_- > 0$. None of these approximations is a consequence of Proposition 1.

$$\mathbb{E}[\Delta\mathrm{MRR}_s] \approx \alpha\beta_+ q_s\eta_s(1-\rho_s) - \alpha\beta_- c_s \approx \alpha\left[\beta_+ q_s\eta_s(1-\rho_s) - \beta_-\kappa\rho_s\right]. \tag{19}$$

In a recurrent-query-limited regime, $q_s \approx \rho_s$ and η_s can be absorbed into the fitted benefit scale, so

$$\mathbb{E}[\Delta\mathrm{MRR}_s] \approx \alpha\beta_+\rho_s(\rho^\star_{\mathrm{rec}} - \rho_s), \qquad \rho^\star_{\mathrm{rec}} = 1 - \frac{\beta_-\kappa}{\beta_+}. \tag{20}$$

In a historical-negative TLP regime, lowering the scores of historical negatives remains useful even for novel positives, so $q_s \approx 1$ and η_s remains an explicit candidate-density correction:

$$\mathbb{E}[\Delta\mathrm{MRR}_s] \approx \alpha\left[\beta_+\eta_s(1-\rho_s) - \beta_-\kappa\rho_s\right], \qquad \rho^\star_{\mathrm{hist}} = \frac{\beta_+\eta_s}{\beta_+\eta_s + \beta_-\kappa}. \tag{21}$$

For illustration, tgbl-wiki has $c \approx 0.409$ and $\rho = 0.709$, giving $\kappa \approx 0.58$. In the historical-negative regime with $q_s = \eta_s = 1$, imposing $\rho^\star_{\mathrm{hist}} = 0.10$ gives $\beta_-/\beta_+ = (1-0.10)/(0.10\kappa) \approx 15.5$. This is a fitted consistency calculation, not an estimated universal threshold; graph support, objective, architecture, and candidate protocol can all alter the effective terms.

G.6 Post-Hoc Recombination of Observed Cells

Table 55 records the post-hoc recombination of observed cells, including the outcome-informed recurrence grouping.

Table 55: Post-hoc recombination of observed cells (recombined non-factorial legacy runs). The ρ row uses an outcome-informed 10% grouping and is not a prospective or deployable policy evaluation; the strict factorial is Table 2.

Policy	tgbl-review	tgbl-comment	tgbl-wiki
Always joint random	0.328 ± 0.018	0.670 ± 0.005	0.788 ± 0.005
Always PAS	0.474 ± 0.007	0.853 ± 0.004	0.661 ± 0.006
Post-hoc $\rho < 10\%$ group	0.474 ± 0.007	0.853 ± 0.004	0.788 ± 0.005

This arithmetic recombination shows what the post-hoc grouping selects among already observed cells. It provides no forward validation of a threshold, and the training-prefix estimate has not been evaluated as a deployment policy.

H Resources and Tuning

Tables 56–57 give the separate DTLP resource comparison and model-specific tuning budgets. Probe defaults are in Appendix C.2; the PDRE recipe is in Appendix E.2. The 11K-parameter and OOM results concern DTLP, not strict PAS.

PAS accounting. Across three co-scheduled seeds, paired RH-versus-RR wall-time changes are $+2.79\%$ on review, $+1.02\%$ on wiki, and -0.06% on comment; these are descriptive timings rather than exclusive-device benchmarks. Routing changes no parameter count, while the explicit pair branch adds 45,668 parameters (182,672 FP32 bytes). Strict-PAS peak VRAM was not recorded. Exact history stores pair counts, pair last-interaction times, and per-source destination sets, requiring state linear in the number of observed directed pairs and therefore unbounded on an unbounded stream. For a fixed node set, a per-source cap bounds the number of stored entries; window expiry alone also depends on arrival volume and node cardinality. These are untested approximations.

Table 56: DTLP resource comparison on thgl-github (5.9M nodes, 17.5M edges; one RTX 4090, 24GB). OOM occurs before the first training step; evaluation was not run.

Method	**Params**	**Train**	**Eval (2.6M queries)**	**GPU mem**
TGN	$\sim$600K	OOM	not run	OOM ($>$ 24GB)
STHN	$\sim$500K	OOM	not run	OOM ($>$ 24GB)
DyGFormer	$\sim$300K	OOM	not run	OOM ($>$ 24GB)
DTLP	11K	3.6 min	3.6 hours	$<$ 2GB

Table 57: Tuning budget per model family.

Model	**Search and settings**	**Runs and configuration basis**
DTLP default	loss $\{$BPR, CE, margin$\}$ $\times$ learning rate $\{5\times10^{-3}, 10^{-4}\}$	6 loss/lr combinations; sampler-inclusive total unverified
TGAT/GraphMixer	no tuned parameters; DyGLib defaults	fixed default settings
DyGFormer	no tuned parameters; DyGLib defaults	fixed default settings
CRAFT	no tuned parameters; published configuration	reported configuration
PDRE residual	$\alpha_{\text{hard}} \in \{0.2, 0.5\}$; collision percentage 75	3–5 runs
PDRE gate	$\lambda_{\text{L1}} = 0.02$; gate learning rate 10^{-3}	matched evaluations
Strict factorial	assignments RR, RH, HR, HH; matched otherwise	3 paired development seeds; 5-seed TGAT confirmation