RACE: RELATION-LEVEL COUNTERFACTUAL EXPLANATIONS FOR HETEROGENEOUS GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual explanations of graph neural networks identify edge deletions that flip a prediction. On heterogeneous graphs, however, existing methods first collapse the graph into untyped edges, so they cannot answer the question a domain expert actually asks: *which relation type* drives this prediction? We present RACE (**R**elation-**A**ware **C**ounterfactual **E**xplanations), which gives this question an exact, per-instance answer. For every explained instance, an exhaustive search over relation subsets returns the certified minimum relation-deletion set that flips the prediction—or an explicit report that no such deletion exists; each relation-level answer is then refined into a typed edge set within the attributed relations, verified on the discrete model by single-edge restoration. The relation-level answer is exact and deterministic given the frozen backbone, whereas soft-mask baselines vary by 6–8 pp in success rate across runs differing only in random ordering. On ACM, a Cora-derived graph, and ogbn-mag, RACE improves counterfactual success rate over the strongest baseline by up to $+2.7$ pp while deleting fewer edges, and attains the highest success rate among all same-task baselines on every dataset; the advantage reproduces across four backbones on ogbn-arXiv and on DBLP, with cross-seed relation-set agreement up to 0.89. A synthetic study with known generating mechanisms confirms that the search recovers the relation the trained model actually relies on—and reports infeasibility rather than fabricating an attribution when the model has learned none—so the explanations stay trustworthy exactly where explanations matter.

1 INTRODUCTION

Graph neural networks (GNNs) are the workhorse for learning on graph-structured data, yet their predictions remain opaque. *Counterfactual* explanations address opacity by identifying the minimal input perturbation that *flips* the prediction: CF-GNNExplainer Lucic et al. (2022), CF^2 Tan et al. (2022), and RCExplainer Bajaj et al. (2021) delete edges until the prediction changes, and recent work keeps extending this line Zhang et al. (2026). Real-world graphs, however, are rarely homogeneous: academic networks, knowledge graphs, and e-commerce platforms are heterogeneous information networks in which co-authorship, citation, and shared-subject links coexist Shi et al. (2017). On such data the existing methods have two defects. First, they *collapse* the heterogeneous graph into a pairwise graph, discarding relation-type semantics, so the resulting explanation cannot answer "*which relation type* drives the prediction"—the question a domain expert asks of a heterogeneous model. Second, even in their native regime, their edge-level outputs are a scattered set of individual edges whose semantics are not aggregable into a human-readable statement such as "co-authorship is what this prediction relies on".

We close this gap with RACE, a framework that (i) preserves relation-type semantics in the explained model, (ii) computes *instance-level* minimal relation-deletion sets by exact search, (iii) refines them into *verified*, single-edge-restoration-irreducible edge deletions inside the attributed relations (a relation-to-edge hierarchy), and (iv) evaluates everything with deletion- and sufficiency-fidelity metrics plus explicit coverage, cost, and stability reporting. We make three contributions:

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.

- **Instance-level relation counterfactuals by exact search.** For each explained instance v we solve $S_v^* = \arg\min_S C_v(S)$ subject to the prediction flipping (a margin constraint), with a lexicographic cost (number of relations, then local edge fraction, then margin). Because K is small, enumeration over the 2^K subsets is exact, deterministic, and cheap ($\leq$10 s at K=10); every instance receives either this certified answer or an explicit infeasibility report—no force-attribution—and the resulting coverage is disclosed. A budgeted branch-and-bound search extends the same contract to larger K (RQ6). We further document *non-monotonicity*: some predictions flip under a relation subset but not under full removal, evidence that aggregate flip-rate searches cannot substitute for the instance-level problem.
- **A verified hierarchical relation-to-edge pipeline.** The edge-level stage deletes the attributed relations and then *verifies* the intervention on the discrete model by restoring edges one at a time along the saliency order while the flip persists, repeated to a fixpoint—so the reported set carries a single-edge-restoration irreducibility certificate (or an explicit budget-limited status), evaluated on the instance's exact local computation subgraph. Under a strict same-frozen-typed-backbone, instance-targeted protocol, the hierarchical explainer beats the strongest baseline (CF^2-hetero) on both success rate ($+0.8$/$+2.7$/$+1.4$ pp on ACM/Cora/ogbn-mag, $p = 0.002$ raw, Holm-corrected ≤ 0.062, ten seeds) and edge cost (-1.9/-1.0/-2.7 pp), reproduces on ogbn-arXiv across four backbones and on DBLP, yields cross-seed relation-set agreement of 0.78–0.89 on the separable datasets, and attains the highest success rate among the same-task counterfactual baselines under the same frozen checkpoint.
- **A fidelity framework and an identifiability characterization.** We adopt the deletion/sufficiency statistics as descriptive fidelity metrics—Necessity Fidelity (NF) and Sufficiency Fidelity (SF)—reserving causal language for settings with an explicit generating mechanism; we characterize *when* relation-level attribution is identifiable (causally separable relations), and—via a five-mechanism synthetic study separating model-dependence from data-causal recovery—show that the search recovers the relation the trained model actually relies on and reports infeasibility instead of fabricating an attribution when the model has not learned a mechanism.

2 Related Work

Explaining GNNs. GNNExplainer Ying et al. (2019) learns a mask over edges and features that *preserves* the prediction, and PGExplainer Luo et al. (2020) amortizes this into a generator; both are factual and cannot answer counterfactual queries. Gradient attributions—Saliency Baldassarre & Azizpour (2019), GradCAM Pope et al. (2019), Integrated Gradients Sundararajan et al. (2017)—and SubgraphX Yuan et al. (2021) score inputs without a flip guarantee. None provides relation-type granularity.

Counterfactual explanations on graphs. CF-GNNExplainer Lucic et al. (2022) formalizes the counterfactual explanation as a minimal set of edge deletions that flips the prediction; CF^2 Tan et al. (2022) adds factual reasoning and a validity term; RCExplainer Bajaj et al. (2021) uses a hinge objective; MEG Numeroso & Bacciu (2021) searches binary masks with a genetic algorithm; NSEG Cai et al. (2025) optimizes a necessity-plus-sufficiency lower bound; InduCE Verma et al. (2024) amortizes the counterfactual objective into an inductive generator; C2Explainer Ma et al. (2025) makes the counterfactual customizable through prior constraints. All operate on homogeneous pairwise graphs: none attributes an explanation to a *relation type*, because the graph they consume no longer contains types. What this line lacks is not post-hoc evaluation—prior methods do evaluate their final counterfactual on the discrete model when reporting validity—but a *selection* procedure that consults the discrete model at every step, per instance, and reports infeasibility instead of force-attributing—the ingredients RACE contributes.

Probabilities of causation and heterogeneous GNNs. Pearl introduced necessity, sufficiency, and their conjunction as counterfactual interpretations of causation Pearl (1999); Tian & Pearl (2000); Cai et al. (2025) brought this framework to graph explanations. We adopt their evaluation conventions as descriptive fidelity metrics (NF/SF) and reserve causal language for synthetic data with an explicit generating mechanism. HAN Wang et al. (2019), HGT Hu et al. (2020b), and R-

GCN Schlichtkrull et al. (2018) organize heterogeneous message passing; type-aware attribution methods remain factual and post-hoc. Heterogeneous explainability has not been combined with (i) minimal *instance-level* counterfactual interventions, (ii) relation-type-level intervention units with explicit coverage, and (iii) per-instance verification on the discrete model—the three ingredients RACE contributes.

3 METHODOLOGY

3.1 PRELIMINARIES AND PROBLEM FORMULATION

A heterogeneous graph is a tuple $\mathcal{G} = (\mathcal{V}, \{E_r\}_{r=1}^{K}, X)$ with relation edge sets $E_r \subseteq \mathcal{V} \times \mathcal{V}$, K relation types ($K \leq 10$ in our benchmarks), and node features $X \in \mathbb{R}^{N \times d}$. A frozen graph neural classifier f maps $\mathcal{G}$ to class probabilities $f(\mathcal{G}) \in [0,1]^{N \times C}$. For an instance v with prediction $\hat{y}_v = \arg\max_c f(\mathcal{G})_{v,c}$, we write the margin

$$M_v(\mathcal{G}) \;=\; f(\mathcal{G})_{v,\hat{y}_v} - \max_{c \neq \hat{y}_v} f(\mathcal{G})_{v,c}, \tag{1}$$

so a flip means $M_v \leq -\kappa$ for a target confidence $\kappa \geq 0$. We study two intervention granularities: *relation-type-level* (delete every edge of one or more relations, written $\mathcal{G} \backslash S$ for $S \subseteq \{1, \ldots, K\}$) and *edge-level* (delete a typed edge subset, written $\mathcal{G} \odot m$ for per-relation keep-masks $m^r \in [0,1]^{|E_r|}$). Throughout, $N_L(v)$ denotes the backward ball of radius L around v (the receptive field of an L-layer backbone), and $E_L(v)$ the edges of the L-layer computation subgraph of v.

3.2 RELATION-AWARE HETEROGENEOUS BACKBONE

The explained model preserves relation semantics through a simplified heterogeneous graph transformer. For each relation r, source features are projected by a dedicated matrix and edges are scored by additive attention:

$$\begin{aligned} h_v^r &= W_r x_v, \\ e_{uv}^r &= \mathrm{LeakyReLU}\big(a_r^\top \big[W_r x_u \,\|\, W_r x_v\big]\big), \\ \alpha_{uv}^r &= \mathrm{softmax}_{v \in \mathcal{N}_r(u)}(e_{uv}^r), \end{aligned} \tag{2}$$

and the layer update combines relations with learnable relation weights λ_r:

$$x'_u = \mathrm{ReLU}\Big(\mathrm{LayerNorm}\Big(W_{\mathrm{self}}\, x_u + \sum_{r=1}^{K} \lambda_r \sum_{v \in \mathcal{N}_r(u)} \alpha_{uv}^r\, W_r x_v\Big)\Big). \tag{3}$$

Stacking L such layers and a linear head yields f. On ACM this backbone attains 0.560 ± 0.007 accuracy, $+3.4$ pp over a collapsed pairwise backbone trained identically; HAN, HGT, and the RACE backbone tie at 0.56–0.57 under the shared protocol, while residual-free GCN/GAT attain 0.23—the accuracy edge is a property of the typed residual design.

3.3 EDGE-LEVEL VERIFIED COUNTERFACTUAL

Per-relation typed masks. For each relation r we learn a keep-mask $m^r \in [0,1]^{|E_r|}$ (Gumbel-Sigmoid relaxation Jang et al. (2017); Maddison et al. (2017)); gradients flow through the relation-specific projections, so interventions remain typed.

Margin objective. With the backbone frozen and the targets restricted to the instances being explained, we minimize

$$\mathcal{L} = \underbrace{\mathbb{E}_v[\max(0,\, M_v(\mathcal{G} \odot m) + \kappa)]}_{\text{flip term}} + \underbrace{\lambda_{sp} \tfrac{1}{E} \textstyle\sum_{r,e}(1 - m_e^r)}_{\text{minimality}} + \underbrace{\beta \tfrac{1}{E} \textstyle\sum_{r,e} m_e^r (1 - m_e^r)}_{\text{binary pressure}}, \tag{4}$$

replacing the soft-probability objective used by prior work: the hinge exerts no gradient on already-flipped instances, concentrating optimization on those near the boundary.

Discrete verification and backward pruning. The relaxed masks are *not* the explanation. We (1) order every edge by its learned keep-score and greedily delete in batches, evaluating the *discrete* model after each batch and keeping the deletion point that maximizes validity minus λ_{sp}-weighted cost; (2) backward-prune: restore deleted edges in descending keep-score while the satisfied set is preserved. The reported masks are therefore verified on the discrete model: unlike a threshold applied once after optimization, the selection itself consults the discrete model at every batch, so the flip status of the reported intervention is a measured property of that intervention rather than an implication of the relaxation.

3.4 Instance-Level Relation Deletion Search

For each explained instance v, the relation-level counterfactual is the minimum-cost relation set whose deletion flips v:

$$S_v^* = \arg \min_{S \subseteq \{1,\dots,K\}} C_v(S) \quad \text{s.t.} \quad M_v(\mathcal{G} \setminus S) \leq -\kappa, \tag{5}$$

with the lexicographic cost

$$C_v(S) = \Big(|S|,\ \tfrac{1}{|E_L(v)|} \textstyle\sum_{r \in S} |E_r \cap E_L(v)|,\ M_v(\mathcal{G} \setminus S) \Big), \tag{6}$$

i.e., fewest relation types first, then the smallest fraction of the instance's receptive-field edges, then the strongest flip. Because K is small, we enumerate all 2^K subsets in increasing size; one forward pass per subset evaluates every instance simultaneously, so the total cost is 2^{K+1} forwards (0.1 s at K=3, 10.3 s at K=10). The procedure is deterministic and returns the exact minimum under the frozen predictor and the stated cost. Instances with no feasible subset are marked infeasible, and the fraction of feasible instances is reported as *coverage*: every instance receives either the certified minimum or an explicit infeasibility report, never a force-attributed relation. For larger vocabularies the budgeted branch-and-bound search of Sec. 4.2 provides the same output contract. Two properties follow from the instance-level formulation and are measured in Sec. 4.2: (i) the output contract—per instance, the certified answer or an explicit infeasibility report, with coverage disclosed (0.17–0.33 on our benchmarks); (ii) non-monotonicity—some instances flip under a subset S but not under full removal, so the aggregate "70% of the full-removal effect" criterion of prior work can both miss and misattribute instances.

3.5 Hierarchical Relation-to-Edge Refinement

For every feasible instance, the relation-level answer is refined into a *certified* typed edge set *inside* S_v^*: starting from $\mathcal{G} \setminus S_v^*$ (already flipped), we restore edges of S_v^* one at a time in descending keep-importance order (gradient saliency of the original prediction), accepting a restoration iff v stays flipped, and repeat full passes until a complete pass changes nothing—the returned set is then *single-edge-restoration irreducible*. If the per-instance forward budget is exhausted first, the answer is reported as a verified flip with the irreducibility check incomplete; the two statuses are reported separately, and a budget-exhausted answer is not marked optimal. The certificate is single-edge-restoration irreducibility rather than global minimality; the gap to the true minimum is measured exactly on enumerable subgraphs (Appendix I). Each check runs on the *exact local computation subgraph* of v—the backward ball of radius L with renumbered nodes, which reproduces the full-graph logits of v exactly—so a check costs a tiny forward instead of a full-graph pass. Instances without a feasible relation set keep the flat verified edge explanation of Sec. 3.3 as a fallback, so overall success is never below the flat method. The output is the hierarchy: *"delete relation S_v^*; within it, these edges."*

3.6 Fidelity Metrics and Overall Pipeline

Following the deletion/sufficiency evaluation convention of Cai et al. (2025) but without claiming causal probabilities, we report per instance and in aggregate:

$$\mathrm{NF}_v(S) = \mathbf{1}[f(\mathcal{G} \setminus S)_v \neq \hat{y}_v], \qquad \mathrm{SF}_v(S) = \mathbf{1}[f(S)_v = \hat{y}_v], \tag{7}$$

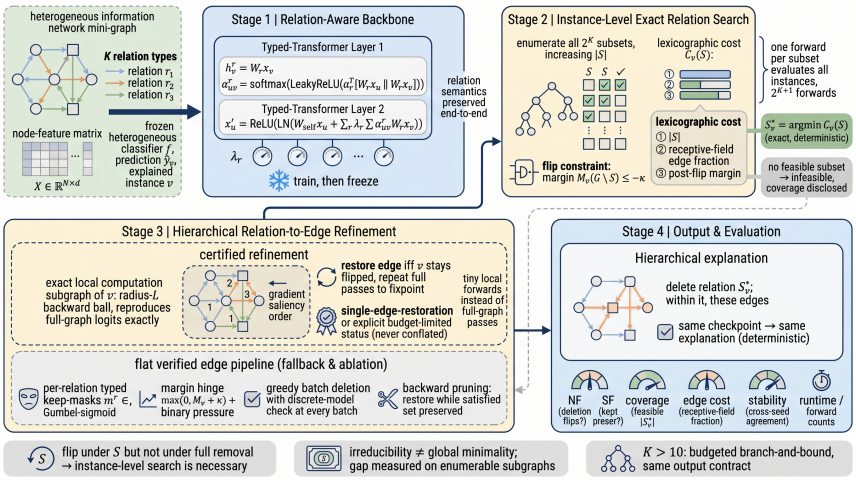


Figure 1: Pipeline of RACE. Stage 1 builds the relation-aware backbone and freezes it. Stage 2 solves the instance-level exact search, returning the certified minimum relation deletion set S_v^* or an explicit infeasibility report. Stage 3 refines each feasible answer into a certified, single-edge-restoration-irreducible typed edge set (or an explicit budget-limited status); relation-infeasible instances keep the flat verified explanation as a fallback. Stage 4 evaluates NF/SF, coverage, cost, stability, and runtime.

Necessity Fidelity (does the deletion flip?) and Sufficiency Fidelity (does the kept set alone preserve the prediction?). Aggregates are instance averages; we additionally report coverage, relation cost $|S_v^*|$, edge cost, the post-intervention margin, cross-seed stability, and runtime, and we never present NF $\times$ SF as a causal joint probability. Fig. 1 depicts the pipeline: **Stage 1** builds the relation-aware backbone (Sec. 3.2); **Stage 2** freezes it and solves the instance-level relation search of Eq. (5), recording S_v^* and feasibility; **Stage 3** refines each feasible answer into a certified, single-edge-restoration-irreducible typed edge set (Sec. 3.5), with infeasible instances keeping the flat verified edge explanation; **Stage 4** evaluates both granularities with NF/SF, coverage, cost, margin, stability, and runtime.

4 EXPERIMENTS

We organize the study around six questions. **RQ1 (main comparison)**: on one frozen typed backbone, does the hierarchical RACE improve over edge-level baselines on validity, cost, and stability? **RQ2 (mechanism)**: is the relation phase the source of the advantage, and what does discrete verification add? **RQ3 (output contract)**: does the instance-level search return the minimum valid relation deletion, with feasibility reported? **RQ4 (generality)**: does the advantage transfer across backbones and datasets? **RQ5 (determinism and robustness)**: how deterministic, stable, and perturbation-robust are the explanations, and what does verification cost? **RQ6 (scalability)**: how do exact and approximate search trade quality against cost as K grows?

4.1 EXPERIMENTAL SETUP

Datasets. *ACM (HAN version)*: a paper network whose relations are meta-path-derived—relation 0: PAP (co-authorship), relation 1: PTP (shared keywords)—1,903-dim bag-of-words features, venue labels. *ogbn-mag (subsample)* Hu et al. (2020a): 30,000 papers of the top-20 venues, relations PAP and citation. *Cora-derived*: citation plus a common-neighbor relation. *ogbn-arXiv (subsample)* Hu et al. (2020a): 30,000 papers of the top-20 subject areas, relations direct citation and co-citation. *DBLP*: an author network (co-authorship plus a shared-term meta-path relation; 4,057 authors, 4 areas). *Synthetic SCM*: a five-mechanism generator (causal, spurious with environment shift, redundant, noise, and a synergistic variant). All four primary real benchmarks have $K = 2$ relation types; a four-relation ogbn-mag variant and the synthetic K-scan probe larger vocabularies. Dataset statistics are summarized in Appendix A.

Table 1: Same-frozen-typed-backbone, instance-targeted comparison (10 seeds, mean±std across seeds; CSR = success rate, cost = mean local deleted fraction over all targets). Paired sign-flip tests of the hierarchical method against each baseline; †: $p = 0.002$ before Holm correction (≤ 0.062 after). **Bold**: best value in each column; underlined: second best.

	Method	ACM CSR↑	ACM cost↓	Cora CSR↑	Cora cost↓	ogbn-mag CSR↑	ogbn-mag cost↓
flat	RACE-per-edge	0.216±0.072	**0.159±0.062**	0.452±0.111	0.305±0.031	0.168±0.008	0.378±0.011
flat	PNS-hetero	0.216±0.072	**0.159±0.063**	0.450±0.110	0.306±0.031	0.168±0.008	**0.378±0.010**
flat	CF^2-hetero	0.216±0.072	0.195±0.088	0.458±0.110	0.342±0.024	0.168±0.008	0.434±0.013
RACE-hier		**0.224±0.077**†	0.176±0.067†	**0.484±0.122**†	0.332±0.024†	**0.182±0.009**†	0.407±0.011†
vs. CF^2-hetero		+0.8 pp	−1.9 pp	+2.7 pp	−1.0 pp	+1.4 pp	−2.7 pp

Metrics. We report backbone accuracy; *counterfactual success rate* (CSR), the fraction of explained instances whose prediction flips, equal to mean NF, with any non-intervention counted as failure; *edge cost*, the fraction of the instance's receptive-field edges deleted, under labeled aggregation conventions; NF/SF of Eq. (7) separately; *coverage*; cross-seed *stability* (relation-set agreement and edge-set Jaccard); and *runtime*. Synthetic studies add two separate hit rates: model-dependence recovery and data-causal recovery. Full definitions are in Appendix C.

Baselines and protocol. One typed backbone is trained per seed and frozen; *every* explainer consumes the same checkpoint, the same explained instances (the correctly-classified test nodes), the same operation (edge deletion), and the same budget; objectives are restricted to the explained instances ("targeted"). Baselines: GNNExplainer-hetero, RACE-per-edge, PNS-hetero, CF^2-hetero, and the NSEG Cai et al. (2025), InduCE Verma et al. (2024), and C2Explainer Ma et al. (2025) objectives, all operating on the typed graph with per-relation masks—the typed counterparts of the published methods, isolating the explanation objective from the architecture. We additionally compare the *flat* verified pipeline (Sec. 3.3 without the relation phase) against the same baselines to isolate the hierarchical contribution (RQ2). HAN Wang et al. (2019), HGT Hu et al. (2020b), and R-GCN Schlichtkrull et al. (2018) serve as the explained models in the cross-backbone study. Intervention semantics are fixed across every evaluation (deletion zeroes masked edge entries before aggregation; the attention softmax is not renormalized; node features and the residual self-projection are never modified). **Statistics.** Ten seeds on ACM/Cora/ogbn-mag (five for cross-backbone and additional-dataset studies); every comparison is a paired test over the same checkpoints and target instances; the primary comparisons are fixed in advance; p-values are exact two-sided sign-flip permutation tests (minimum attainable $p = 0.002$ with ten seeds, 0.0625 with five), supplemented by paired bootstrap 95% CIs and paired Cohen's d; Holm correction is applied across the full family of 12 primary tests. Full hyperparameters and settings are in Appendix B.

4.2 Results

RQ1: main comparison. Table 1 reports the same-frozen-backbone, targeted comparison over ten seeds. At the edge level the flat explainers tie: RACE-per-edge and PNS-hetero are statistically identical on all three datasets ($p \geq 0.38$), and the multi-term CF^2-hetero is at most $+0.1$ pp higher while deleting substantially more edges. The *hierarchical* method breaks the tie: against the strongest baseline (CF^2-hetero) it improves CSR by $+0.8$/$+2.7$/$+1.4$ pp ($p = 0.002$; Holm-corrected ≤ 0.062 across the full test family) while deleting -1.9/-1.0/-2.7 pp *fewer* edges ($p = 0.002$), and its CSR gains over RACE-per-edge and PNS-hetero are significant on all three datasets ($p = 0.002$). Prior methods trade success rate against deletion cost; the hierarchy improves both at once. Stability follows the same direction: pairwise cross-seed relation-set agreement is 0.779 ± 0.058/0.641 ± 0.096/0.894 ± 0.020 (ACM/Cora/mag), and the certified edge-set Jaccard is 0.436 ± 0.041/0.483 ± 0.072/0.654 ± 0.032—the separable datasets show the highest relation-set agreement; on Cora the relation sets vary more across seeds, an identifiability effect analyzed under RQ3 and Sec. 4.2.

Comprehensive comparison. Table 2 compares all same-task counterfactual methods on the same frozen checkpoints with per-method RNG control. RACE-hier attains the highest CSR on every dataset (0.224/0.484/0.182), ahead of the strongest baseline by $+0.8$/$+2.6$/$+1.4$ pp. NSEG's necessity-sufficiency objective essentially never flips (CSR ≤ 0.01), confirming that flip success

Table 2: Same-task comprehensive comparison under one frozen checkpoint family (10 seeds, mean±std). Cost annotations are restricted to the CSR columns, because a method that rarely flips can trivially attain low cost; the paper's cost claim is the paired comparison against CF^2-hetero (Table 1). RACE-hier additionally certifies every reported flip, provides relation-type attribution, and reports infeasibility per instance. **Bold**: best value; underlined: second best.

Method	ACM		Cora		ogbn-mag	
	CSR↑	cost↓	CSR↑	cost↓	CSR↑	cost↓
counterfactual						
CF-GNNExplainer	0.216±0.072	0.159±0.063	0.452±0.110	0.305±0.031	0.168±0.008	0.379±0.011
RC	0.148±0.049	0.145±0.059	0.400±0.100	0.245±0.013	0.088±0.007	0.314±0.012
NSEG	0.000±0.001	0.139±0.041	0.001±0.001	0.224±0.021	0.002±0.001	0.192±0.014
CF^2	0.216±0.072	0.195±0.088	0.458±0.110	0.342±0.024	0.168±0.008	0.434±0.013
C2	0.190±0.069	0.406±0.023	0.448±0.113	0.518±0.024	0.123±0.007	0.275±0.013
INDUCE	0.099±0.091	0.449±0.438	0.195±0.052	0.673±0.348	0.083±0.069	0.519±0.438
search						
MEG	0.072±0.017	0.342±0.004	0.115±0.030	0.342±0.010	0.061±0.006	0.348±0.007
RACE-hier (ours)	**0.224±0.077**	0.176±0.067	**0.484±0.122**	0.332±0.024	**0.182±0.009**	0.407±0.011

is not what it optimizes; INDUCE is unstable at this scale (std up to 0.44); C2 pays the highest deletion cost among the counterfactual objectives on ACM and Cora. The factual, gradient, and random diagnostics are reported separately in Appendix D: factual objectives do not optimize flips by construction, and the gradient attributions use an instance-blind global budget, so they are not same-task comparisons and are excluded from win claims. The comparison also isolates a property that only the exact search offers: the soft baselines' CSR varies substantially across runs that differ only in random ordering (CF^2: 0.216 vs. 0.276 on ACM, 0.452 vs. 0.687 on Cora), while RACE's relation-level answer is deterministic given the frozen backbone. On the nodes where both CF^2 and RACE-hier flip, the hierarchical edge set is cheaper in 66%/44%/65% of cases (ACM/Cora/ogbn-mag), and RACE-hier flips 109/85/394 more nodes than CF^2 while CF^2 flips none that RACE-hier misses (Pareto frontier in Fig. 2, Appendix D).

RQ2: the relation phase is the mechanism. Two controlled comparisons isolate where the advantage comes from. *Flat verified pipeline.* Removing only the relation phase (greedy saliency deletion plus the same single-edge scan and budget, no relation search) leaves CSR at 0.174/0.246/0.165 on ACM/Cora/ogbn-mag—below the strongest baseline and far below the hierarchy—at a lower edge cost (Table 3); adding the relation phase converts this deficit into the significant CSR gains of Table 1. *Relation-only* explanations (no edge refinement) reach only the coverage bound (0.17–0.34, Table 4); the hierarchy strictly dominates both ends. *Where the success comes from.* Table 8 decomposes the ten-seed success rate: on ACM the relation phase contributes 18.2 of the 22.4 pp CSR at a mean cost of 0.105—half the strongest baseline's cost (0.195); on ogbn-mag it delivers 16.7 of 18.2 pp at 0.273 vs. 0.434; on Cora it still delivers 33.7 of 48.4 pp at 0.311 vs. 0.342. *Relation switch.* Table 3 runs the two pipelines under identical machinery and budgets—the only difference is the relation phase. The switch attributes +5.0/+23.8/+1.7 pp of CSR to the relation phase on ACM/Cora/ogbn-mag (all $p = 0.002$) and +5.2/+5.7 pp on ogbn-arXiv/DBLP ($p = 0.0625$, five seeds). On the nodes where both flip, the flat path deletes fewer edges (e.g., ACM 0.033 vs. 0.107), so the relation phase buys the significant CSR gain and the relation-type attribution at a higher per-explanation edge cost, while the pipeline-level cost advantage over CF^2-hetero is the separate, all-targets claim of Table 1. *Verification.* Every reported flip is certified on the discrete model; the scan costs 1.5–6 min per seed against 2–5 s for the soft-mask baselines—an inference-only cost that buys a per-instance certificate, attained for 35–83% of the successful explanations (the remainder are reported as verified with the irreducibility check budget-limited). Budget curves and component ablations are in Appendix E.

RQ3: exactness and the output contract. Table 4 reports the instance-level relation search. Coverage is 0.18 (ACM), 0.34 (Cora), and 0.17 (ogbn-mag): for most correctly-classified instances *no* relation deletion flips the prediction, and RACE says so per instance instead of force-attributing a relation. Where a deletion exists, it is cheap in relation count (1.07–1.14 relations on average) and strong (mean post-flip margin −0.20 to −0.46), with sufficiency fidelity 0.86–0.96. The attribution is dataset-consistent: PAP on ACM (flip 0.150 vs. 0.033), citation on ogbn-mag (0.130 vs.

Table 3: Relation switch (same machinery, scan budget, and target nodes; the only difference is the relation phase). CSR and feasible-success cost, mean±std over seeds; p: exact paired sign-flip. The relation phase buys a significant CSR gain at a higher per-explanation edge cost; the pipeline-level cost advantage over CF^2-hetero is the separate claim of Table 1.

Dataset	hier CSR↑	flat CSR	p	hier cost↓	flat cost↓
ACM	**0.224±0.077**	0.174±0.046	0.002	0.105	**0.033**
Cora	**0.484±0.122**	0.246±0.071	0.002	0.311	**0.080**
ogbn-mag	**0.182±0.009**	0.165±0.009	0.002	0.273	**0.220**
ogbn-arXiv	**0.283±0.005**	0.230±0.003	0.0625	0.287	**0.139**
DBLP	**0.161±0.035**	0.104±0.013	0.0625	0.157	**0.030**

Table 4: Instance-level relation search (10 seeds, mean±std). Coverage = fraction of correctly-classified test instances with any feasible relation deletion.

	ACM	Cora	ogbn-mag
coverage ↑	0.182±0.048	0.337±0.096	0.167±0.009
relation cost ↓	1.07±0.03	1.14±0.02	1.07±0.01
edge cost ↓	0.574±0.013	0.605±0.044	0.838±0.014
margin	−0.263±0.033	−0.459±0.042	−0.197±0.015
SF ↑	0.918±0.031	0.858±0.028	0.959±0.009
attribution	PAP (10/10)	common-nbr (10/10)	citation (10/10)

0.032), and common-neighbor on Cora (0.208 vs. 0.130, a consistent winner across all ten seeds); it also varies *by class*, recovering per-class relation semantics that a dataset-level rate cannot (Appendix F). Non-monotonicity is measurable: on ACM, coverage (0.18) *exceeds* the full-removal flip rate (0.14)—instances flip under a subset but not under full deletion, evidence that dataset-level flip-rate searches cannot substitute for the instance-level problem. Four archetypes recur across the explained nodes (single-relation flip, two-relation flip, explicit refusal, flat-fallback success; Appendix F), and bucketing by prediction margin shows the hierarchical CSR is highest exactly where predictions are closest to the decision boundary.

RQ4: generality across backbones and datasets. Table 5 reports ogbn-arXiv under four frozen backbones (five seeds each): the hierarchical CSR advantage over the strongest baseline reproduces on $4/4$ backbones ($+0.9$ to $+2.6$ pp, all $p = 0.062$), and the cost advantage over CF^2-hetero on $4/4$ (-4.1 to -7.1 pp); R-GCN, whose simpler aggregation makes more predictions relation-deletable, yields the highest CSR (0.490). The backbone study extends to ACM, Cora-derived, and ogbn-mag (three backbones each, five seeds; Table 10, Appendix G): the hierarchical method beats CF^2-hetero on CSR in all $9/9$ cells ($+1.0$ to $+8.1$ pp, all $p = 0.062$). DBLP reproduces the CSR advantage—0.161 vs. 0.145 for CF^2-hetero ($+1.6$ pp, five seeds, $p = 0.0625$). *Natural multi-relation vocabulary.* A four-relation variant of ogbn-mag (PAP, citation, shared-field-of-study, shared-institution) raises coverage from 0.17 to 0.27, and the hierarchy still beats CF^2-hetero (0.322 vs. 0.299, $+2.3$ pp, ten seeds, $p = 0.002$) and the flat switch ($+9.3$ pp, $p = 0.002$), with the attribution staying on citation (5/5 seeds). Taken with Table 1, the mechanism is dataset- and backbone-agnostic, and the relation-level contract is not an artifact of $K{=}2$ vocabularies.

RQ5: determinism and robustness. Given the frozen backbone, the relation-level answer is exact and deterministic; across seeds the relation sets agree at 0.89/0.62/0.89 (ACM/Cora/mag), and under 5% random edge add/remove the attributed relation sets remain near-identical (S_v^* agreement 0.975±0.016/0.930±0.023/0.965±0.011). Sensitivity ablations (Table 11, Appendix H) confirm the reported results are not artifacts of the margin threshold or the cost definition: $\kappa{=}0.05$ lowers coverage by -2.3/-1.0/-3.3 pp (the expected effect of a stricter flip threshold), and switching the cost criterion leaves coverage identical on all three datasets.

RQ6 and the Synthetic SCM: scalability and identifiability. A synthetic five-mechanism generator (CF-GNNExplainer-style Lucic et al. (2022); causal, spurious with environment shift, redundant, noise, and a synergistic variant; generation details in Appendix J) separates *model-dependence* recovery (the returned set equals the relation the trained model relies on) from *data-causal* recovery (it equals the generating relation). In the benchmark regime ($K \leq 10$), exact search certi-

Table 5: ogbn-arXiv: hierarchical vs. strongest baseline per backbone (5 seeds; all CSR/cost comparisons $p = 0.062$, the minimum attainable with five seeds). RACE-bb is the relation-aware backbone of Sec. 3.2. **Bold**: better value in each pairwise comparison.

	CSR↑		cost↓	
Backbone	RACE-hier	CF^2	RACE-hier	CF^2
RACE-bb	**0.283**	0.267	**0.287**	0.385
HAN	**0.255**	0.245	**0.284**	0.379
HGT	**0.288**	0.275	**0.286**	0.430
R-GCN	**0.490**	0.464	**0.315**	0.510

fies the minimum-cost answer at negligible wall-clock cost (0.1 s at $K{=}3$ to 7 s at $K{=}10$, faster than the effect-ordered greedy/beam heuristics for $K \leq 5$), and the double hit-rate protocol reports model-dependence equal to data-causal recovery—the regularized model relies on the generating relation. The heuristics attain higher data-causal hit only because the lexicographic tie-break prefers cheaper noise relations over the causal one when both flip; exact search's guarantee is the certified minimum-cost answer. When the mechanism is not learnable (synergistic variant, test accuracy $0.25 \approx$ chance), all data-causal hits collapse to chance while model-dependence sits at 0.20: the double protocol attributes the collapse to model unlearnability rather than explainer error, which a single hit rate conflates. At $K{=}20$, where 2^{20} enumeration is impractical, the budgeted branch-and-bound search (a first-order Lipschitz bound, capped at 500 local forwards per instance) is the only approximation that improves over greedy/beam (0.16 vs. 0.06 data-causal hit). The full results are in Table 13 (Appendix J); the same certification discipline applies at the edge level, where the scan equals the true minimum in 85% of enumerable cases and is off by one edge in the rest (Appendix I).

5 CONCLUSION

We presented RACE, which explains a frozen heterogeneous GNN at two granularities that prior work conflates: an *instance-level* minimal relation-deletion set found by exact, deterministic search, and a *verified*, single-edge-restoration-irreducible typed edge set inside the attributed relations. Under a strict same-frozen-typed-backbone, instance-targeted protocol, the hierarchical explainer significantly improves success rate while deleting fewer edges than the strongest baseline on ACM, Cora-derived, and ogbn-mag; the advantage reproduces across four backbones on ogbn-arXiv, on DBLP, and on a natural four-relation ogbn-mag; and across the same-task counterfactual baselines, RACE attains the highest success rate on every dataset. Two properties make the answers dependable in practice: determinism—the same checkpoint yields the same explanation, with cross-seed relation-set agreement up to 0.89 and 0.93–0.98 agreement under 5% edge perturbation—and the per-instance output contract—every reported flip is certified on the discrete model, and every infeasible instance is reported as such rather than force-attributed. Future work extends the intervention unit from relation types to meta-paths, grounds relation-level fidelity in explicit structural causal models, scales the certified search past $K{=}10$ with provable approximation bounds, and evaluates hierarchical explanations with human users.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper. We have not used generative AI tools for any task requiring disclosure—generating synthetic data, developing theoretical models or conceptual frameworks, formulating mathematical claims or assisting their proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results; these tasks are not applicable to this work. All research ideas, methods, experiments, analyses, figures, tables, and claims were produced and verified by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All datasets used in this work are public benchmarks (ACM, ogbn-mag, ogbn-arXiv, Cora, DBLP) or synthetic generators described in Sec. 4.2; the subsampling and preprocessing steps are specified in Sec. 4.1 and Appendix B. The complete experimental protocol—hyperparameters, intervention semantics, and statistical procedures—is given in Sec. 4.1 and Appendix B. An anonymized implementation of RACE (backbone training, exact relation search, hierarchical refinement, and the evaluation harness) is available at `https://anonymous.4open.science/r/race_official-68A1`.

REFERENCES

Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. Robust counterfactual explanations on graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 5644–5655, 2021.

Federico Baldassarre and Hossein Azizpour. Explainability techniques for graph convolutional networks. In *ICML 2019 Workshop on Learning and Reasoning with Graph-Structured Representations (LRGSD)*, 2019.

Ruichu Cai, Yuxuan Zhu, Xuexin Chen, Yuan Fang, Min Wu, Jie Qiao, and Zhifeng Hao. On the probability of necessity and sufficiency of explaining graph neural networks: A lower bound optimization approach. *Neural Networks*, 184:107065, 2025.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 22118–22133, 2020a.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the Web Conference 2020 (WWW)*, pp. 2704–2710, 2020b.

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

Ana Lucic, Maartje A. Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pp. 4499–4511. PMLR, 2022.

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

Jiali Ma, Ichigaku Takigawa, and Akihiro Yamamoto. C2explainer: Customizable mask-based counterfactual explanation for graph neural networks. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 137–149, 2025.

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

Danilo Numeroso and Davide Bacciu. Meg: Generating molecular counterfactual explanations for deep graph networks. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999.

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10772–10781, 2019.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference (ESWC)*, pp. 593–607. Springer, 2018.

Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1): 17–37, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 3319–3328. PMLR, 2017.

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pp. 1018–1027, 2022.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Samidha Verma, Burouj Armgaan, Sourav Medya, and Sayan Ranu. Induce: Inductive counterfactual explanations for graph neural networks. *Transactions on Machine Learning Research (TMLR)*, 2024.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *The World Wide Web Conference (WWW)*, pp. 2022–2032, 2019.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 9244–9255, 2019.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 12241–12252. PMLR, 2021.

Yu Zhang, Sean Bin Yang, Arijit Khan, and Cuneyt Gurcan Akcora. Atex-cf: Attack-informed counterfactual explanations for graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2026.

A Dataset Statistics

Table 6 summarizes the datasets used in the experiments.

Table 6: Datasets used in the experiments.

Dataset	#Nodes	#Classes	Features	Relation types
ACM (HAN)	12,499	14	1,903-d BoW	PAP, PTP
ogbn-mag (sub)	30,000	20	128-d emb.	PAP, citation
ogbn-mag (sub, 4-rel)	30,000	20	128-d emb.	PAP, citation, PFP, PAIP
Cora-derived	2,708	7	1,433-d BoW	citation, common-neighbor
ogbn-arXiv (sub)	30,000	20	128-d emb.	citation, co-citation
DBLP	4,057	4	334-d BoW	co-authorship, shared-term
Synthetic SCM	400	4	16-d noise	causal, spurious, redundant, noise $\times(K-3)$

B Extended Experimental Settings

Backbone. Hidden dim 32, 2 layers, dropout 0.5, weight decay 5×10^{-4}, early stopping. **Explainer.** 150 mask steps (Adam, lr 0.1, $\lambda_{sp} = 0.05$, Gumbel $\tau = 1.0$), then discrete verification (40–80 deletion batches) and up to three backward-pruning sweeps. **Relation search.** Exact enumeration, $\kappa = 0$; the margin constraint of Eq. (5) is applied uniformly in the relation search and in the discrete edge-level checks (ties count against the original class). **Intervention semantics.** Deletion zeroes the masked edge entries before aggregation; the attention softmax is *not* renormalized over the surviving edges (it is computed on the intact edge set); node features and the residual self-projection are never modified; edge sets are stored symmetrically so each direction is an independent maskable message entry, and all costs count these directed message entries. Every check runs the frozen model in eval mode with the original predicted class held fixed. **Verification cost.** The restoration scan costs 1.5–6 min per seed against 2–5 s for the soft-mask baselines, with the cost concentrating in the per-node restoration passes (relation enumeration 0.1–0.7 s, one saliency pass 0.01 s). **Statistics.** Aggregates over seeds use the per-seed (macro) mean unless a table caption says otherwise; hierarchical fallback records are imputed at the per-seed CF^2 mean cost in the all-targets cost convention of Table 1.

C Extended Metric Definitions

Backbone accuracy. Test accuracy of f on the intact graph—reported separately from explanation quality (the explained model, not the explainer). **Counterfactual success rate (CSR).** The fraction of explained instances whose prediction flips under the intervention; equals mean NF. A method that returns no intervention (infeasible, or budget exhaustion without a verified flip) counts as a failure, never as a zero-cost success. **Edge cost.** The fraction of the instance's receptive-field edges deleted (reported per instance and averaged); relation cost is $|S_v^*|$. Cost is reported under labeled conventions—over all targets (failures assigned the full-cost penalty 1.0), over each method's successful subset, and over the subset on which all compared methods succeed—and no aggregate mixes them without a label. **NF / SF.** Necessity and sufficiency fidelity of Eq. (7), reported separately. **Coverage.** The fraction of explained instances with a feasible relation-level deletion (for the relation-level and hierarchical methods). **Stability.** Cross-seed agreement of the relation set (identity agreement) and Jaccard of the deleted edge sets; for baselines, Jaccard of their masks across seeds. **Runtime.** Wall-clock per seed, plus forward-pass counts for the search procedures. **Hit rates (synthetic only).** Model-dependence recovery (the returned set equals the relation the trained model actually relies on) and data-causal recovery (the returned set equals the generating relation)—reported separately, since they need not agree.

D Diagnostic Comparison and Pareto Frontier

Table 7 reports the factual, gradient, and random diagnostics on the same checkpoints. Factual objectives (GNNExplainer, PGExplainer) optimize prediction preservation, so near-zero CSR is by construction; the gradient attributions delete a single instance-blind global top-5% edge budget and are not per-instance interventions; Random is a sanity floor. These rows are not same-task baselines and are excluded from win claims. Figure 2 shows the success-rate-vs-cost frontier across all compared methods, and Fig. 3 contrasts the hierarchical method with the strongest baseline on CSR and cost directly.

Table 7: Diagnostic comparison (same checkpoints, 10 seeds, mean±std). Factual objectives optimize prediction preservation, so near-zero CSR is by construction; the gradient attributions delete a single instance-blind global top-5% edge budget and are not per-instance interventions; Random is a sanity floor. These rows are not same-task baselines and are excluded from win claims.

	ACM		Cora		ogbn-mag	
Method	CSR↑	cost↓	CSR↑	cost↓	CSR↑	cost↓
factual						
GNNExplainer	0.000±0.001	0.139±0.041	0.001±0.001	0.224±0.021	0.002±0.001	0.192±0.014
PGExplainer	0.038±0.058	0.250±0.354	0.022±0.048	0.039±0.080	0.056±0.062	0.358±0.412
gradient (global top-5% edge budget, instance-blind)						
Saliency	0.142±0.026	0.103±0.006	0.158±0.037	0.181±0.030	0.150±0.007	0.611±0.012
GradCAM	0.195±0.053	0.100±0.005	0.379±0.080	0.131±0.015	0.162±0.008	0.483±0.012
random						
Random	0.010±0.003	0.050±0.001	0.028±0.010	0.052±0.004	0.009±0.002	0.050±0.003

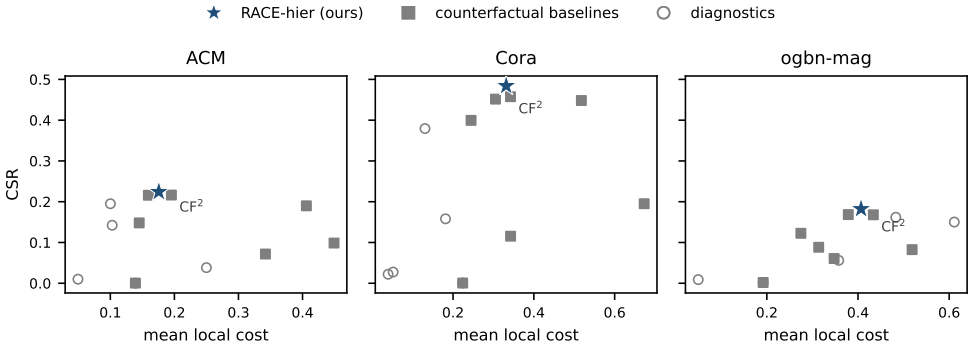


Figure 2: Success rate vs. mean local deletion cost across the compared methods (10 seeds; same-task counterfactual methods in filled markers, diagnostics in open markers). RACE-hier (star) is the highest-CSR point on all three datasets, Pareto-dominating the strongest baseline CF^2 (annotated).

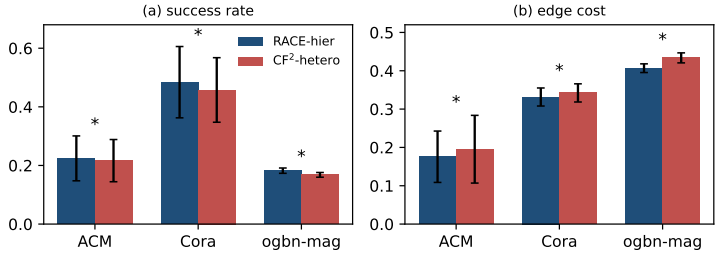


Figure 3: Hierarchical vs. the strongest baseline (CF^2-hetero): (a) counterfactual success rate and (b) mean local edge cost (10 seeds, mean ± 1 std; ∗: paired sign-flip $p = 0.002$, Holm-corrected ≤ 0.062). The hierarchy improves both metrics at once on every dataset.

E Mechanism: CSR Decomposition, Budget Curves, and Component Ablation

CSR decomposition. Table 8 decomposes the ten-seed success rate into the relation-phase answers and the flat fallback (per-seed contributions, so the two phases sum to the total by construction). The flat fallback only extends coverage where no relation deletion exists, so the hierarchy is not a wrapper around the flat explainer but the mechanism that produces the double advantage.

Table 8: Where the success rate comes from (10 seeds): relation-phase answers vs. the flat fallback. The relation phase delivers most of the CSR at roughly half the strongest baseline's cost.

	ACM	Cora	ogbn-mag
relation-phase answers (certified, n)	2546	1106	4185
CSR contribution (pp)	18.2±4.8	33.7±9.6	16.7±0.9
mean edge cost (fraction)	0.105	0.311	0.273
fallback CSR (pp)	4.2±2.9	14.7±2.9	1.5±0.3
total CSR	**0.224**	**0.484**	**0.182**
CF^2 mean cost (ref.)	0.195	0.342	0.434

Component ablation. Table 9 isolates each component on the same checkpoints. Removing the flat fallback drops CSR by exactly the fallback's contribution ($-4.2/-14.7/-1.5$ pp, matching Table 8); the scan budget trades cost against compute along the budget curves (coarser budgets are faster but coarser; the default budget 128 is the working point used throughout). The "full" row reproduces the numbers of Table 1, as the pipeline's core is deterministic given the frozen checkpoint; the rows are derived from the per-node scan traces, so no separate runs were needed.

Table 9: Component ablation of the hierarchical pipeline (10 seeds, mean±std; CSR over all targets, cost = mean feasible edge cost at the given scan budget). full = saliency prior + CF^2 fallback + single-edge scan (budget 128).

Variant	ACM		Cora		ogbn-mag	
	CSR↑	cost↓	CSR↑	cost↓	CSR↑	cost↓
full (budget 128)	0.224±0.077	0.107	0.484±0.122	0.313	0.182±0.009	0.273
− fallback	0.182±0.048	0.107	0.337±0.096	0.313	0.167±0.009	0.273
budget 64	0.182	0.176	0.322	0.390	0.167	0.273
budget 32	0.182	0.279	0.322	0.458	0.167	0.287

Budget curves. Reconstructing the per-node scan traces at forward budgets $\{8, \ldots, 128\}$ (Fig. 4) shows the hierarchical cost decreasing monotonically with budget (ACM $0.490 \to 0.107$; Cora $0.552 \to 0.313$) at unchanged CSR, while the flat variant saturates after ≤ 16 forwards; the curves give the common working point at any budget.

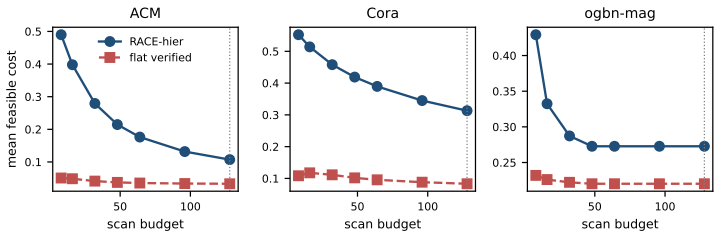


Figure 4: Budget curves (10 seeds): mean feasible edge cost vs. the per-node scan budget for the hierarchical (solid) and flat-verified (dashed) pipelines, reconstructed from the per-node scan traces. The hierarchy's cost decreases monotonically with budget at unchanged CSR; the flat variant saturates after ≤ 16 forwards. The default working point is budget 128.

F Output Contract: Attribution Detail, Archetypes, and Explanation Structure

Per-class attribution. On ogbn-mag, classes 1/2/6/11/13/16/17/18 are citation-dominated (68–100% of relation-attributed nodes) while classes 0/3/4/14/19 lean on co-authorship (31–60%); on ACM, co-authorship dominates every venue class (62–86%) with the keyword relation contributing 6–30%. The instance-level search thus recovers per-class relation semantics that a dataset-level rate cannot. Figure 5 reports the single-relation flip rates behind these attributions.

Archetypes. Four archetypes recur across the explained nodes (ACM, seed 0; Fig. 7): a single-relation flip (node 3: delete co-authorship, $1/87$ local edges, margin -0.21, irreducible certificate), a two-relation flip (node 6: co-authorship plus keyword, $8/91$ local edges, margin -0.032, irreducible), an explicit refusal (node 18: no relation deletion flips, margin 0.83), and a flat-fallback success (node 76). The same pipeline returns a human-readable relation answer where one exists, and says so where it does not.

Explanation structure. Figure 6 summarizes the structure of the explanations: the certified edge sets are cheap (mean local cost 0.11–0.31); a single relation suffices for 86–93% of feasible instances; and the final margins separate flipped from refused instances. *Margin buckets.* Bucketing the explained nodes by their prediction margin (10 deciles, 10 seeds, pooled), the hierarchical CSR is highest in the lowest-margin decile and falls monotonically across deciles on all three datasets (ACM 0.693 to 0.015; Cora 0.885 to 0.177; ogbn-mag 0.396 to 0.030): the method delivers relation-level counterfactuals exactly where predictions are closest to the decision boundary—the instances for which an explanation is most consequential.

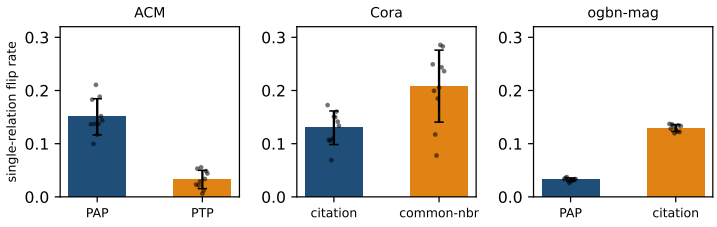


Figure 5: Single-relation flip rates (10 seeds, mean $\pm$ 1 std, dots: per-seed values): ACM (separable: co-authorship PAP dominates), Cora (common-neighbor wins with moderate separation), ogbn-mag (citation dominates).

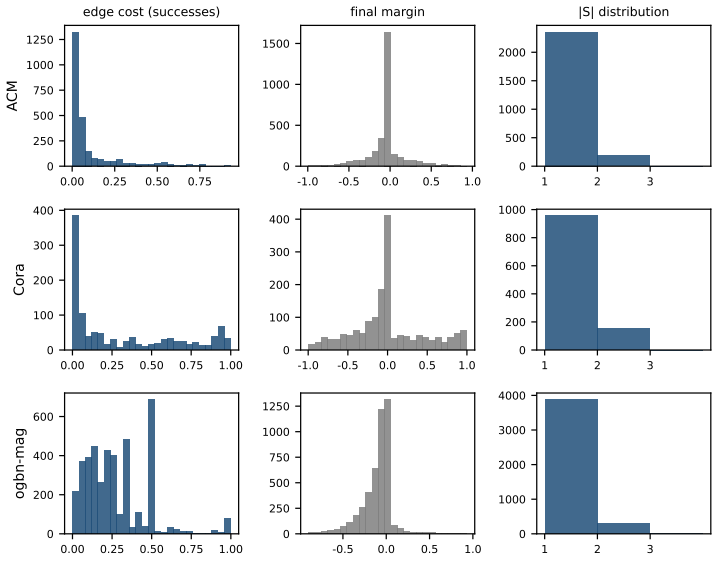


Figure 6: Structure of the explanations (10 seeds): the certified edge sets are cheap (mean local cost 0.11–0.31); a single relation suffices for 86–93% of feasible instances; and the final margins separate flipped from refused instances.

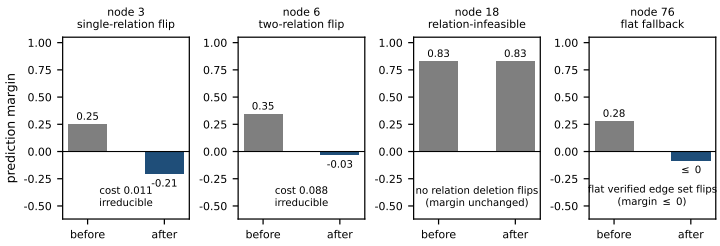


Figure 7: Case studies (ACM, seed 0): the four archetypes of the output contract. (a) a single-relation flip; (b) a two-relation flip; (c) an explicit refusal—no relation deletion flips the prediction, reported as infeasible; (d) a relation-infeasible node that the flat verified edge set flips instead. Bars show the prediction margin before/after the intervention (hatched: no relation-level change exists); edge costs are the deleted fraction of the node's receptive field.

G GENERALITY: CROSS-BACKBONE STUDY

Table 10 extends the backbone study of Table 5 to ACM, Cora-derived, and ogbn-mag (three backbones each, five seeds): the hierarchical method beats CF^2-hetero on CSR in all $9/9$ backbone–dataset cells ($+1.0$ to $+8.1$ pp, all $p = 0.062$, the minimum attainable with five seeds). The paired cost advantage is established in the all-targets comparison of Table 1 and on ogbn-arXiv (Table 5).

Table 10: Cross-backbone study on ACM / Cora-derived / ogbn-mag (5 seeds; hierarchical vs. CF^2-hetero, CSR = success rate; all CSR differences $p = 0.062$, the minimum attainable with five seeds). The CSR advantage holds in all $9/9$ backbone–dataset cells; the paired cost advantage is established in the all-targets comparison of Table 1 and on ogbn-arXiv (Table 5). **Bold**: better value in each pairwise comparison.

Backbone	Method	ACM CSR↑	Cora CSR↑	ogbn-mag CSR↑
HAN	RACE-hier	**0.255**	**0.314**	**0.182**
	CF^2-hetero	0.244	0.298	0.164
HGT	RACE-hier	**0.384**	**0.611**	**0.205**
	CF^2-hetero	0.355	0.581	0.185
R-GCN	RACE-hier	**0.785**	**0.939**	**0.318**
	CF^2-hetero	0.771	0.858	0.292

H Robustness and Sensitivity Ablations

Table 11 reports the κ-sensitivity and cost-criterion ablations plus relation-set robustness under 5% random edge add/remove: κ=0.05 lowers coverage by $-2.3/-1.0/-3.3$ pp (the expected effect of a stricter flip threshold, ACM/Cora/mag); switching the cost criterion leaves coverage identical on all three datasets ($0.1818/0.3370/0.1674$); and the attributed relation sets remain near-identical under perturbation.

Table 11: Sensitivity ablations (ACM/Cora/mag, 10 seeds): margin threshold and cost criterion, plus relation-set robustness under 5% random edge add/remove.

Ablation	Setting	Result (ACM / Cora / ogbn-mag)
margin threshold κ	κ=0.05 vs. κ=0; coverage	0.159 ± 0.044 vs. 0.182; 0.327 ± 0.095 vs. 0.337; 0.135 ± 0.009 vs. 0.167
cost criterion	type / edge / lexicographic; coverage	identical on all datasets (0.1818 / 0.3370 / 0.1674)
edge perturbation	5% random add/remove; S_v^* agreement	0.975 ± 0.016 / 0.930 ± 0.023 / 0.965 ± 0.011

I Edge-Level Certification on Enumerable Subgraphs

On enumerable computation subgraphs (synthetic SCM, 1-layer backbone so radius-1 computation subgraphs are enumerable; 120 instances) we certify the restoration scan against full edge enumeration (Table 12): the certified heuristic equals the true minimum deletion count in $102/120$ cases and is off by one edge in the five other feasible cases; non-monotonicity (a flip that disappears when one edge is added back) occurs in $60/120$ instances.

Table 12: Exact edge-level certification on enumerable subgraphs (120 instances). The certified heuristic equals the true minimum deletion count in $102/120$ cases and is off by one edge in the five other feasible cases.

Quantity	Value	Note
scan = true minimum	$102/120$ (85%)	scan vs. full enum.
off by exactly one edge	5	only nonzero gaps
non-monotonic instances	$60/120$ (50%)	flip under S, not $S \cup \{e\}$

J SYNTHETIC SCM: ADDITIONAL DETAIL

Generation. Features are iid noise; labels are determined by the mean of the *causal*-relation neighbors' features projected onto class patterns; a *spurious* relation is homophilic in an attribute correlated with the label in the train environment but independent in the test environment; a *redundant* relation duplicates the causal signal; the rest are Erdős–Rényi noise; a *synergistic* variant splits the label into two parts recoverable only from two relations jointly. For each configuration and ten data seeds we train the frozen backbone and report the two hit rates of Sec. 4.2. Table 13 reports the full K-scan: exact search attains $0.529/0.292/0.151$ data-causal hit at $K{=}3/5/10$ while the effect-ordered heuristics attain $0.634/0.440/0.286$.

Table 13: Synthetic SCM (10 data seeds; hit = data-causal recovery; model-dependence recovery coincides with it at $K{\leq}10$ and is 0.089 at $K{=}20$ / 0.201 synergistic; exact at $K \leq 10$, greedy/beam everywhere, bnb at $K{=}20$). Exact search certifies the minimum-cost answer and is the fastest search in the benchmark regime ($K \leq 10$); greedy/beam are effect-ordered heuristics, and at $K{=}20$ the budgeted branch-and-bound (bnb) is the only approximation that improves over them.

Config	exact	greedy	beam	bnb	time (ex./gr.)
standard $K{=}3$	0.529	0.634	0.633	—	0.07 s / 1.0 s
standard $K{=}5$	0.292	0.440	0.438	—	0.23 s / 2.5 s
standard $K{=}10$	0.151	0.286	0.285	—	7.0 s / 5.6 s
standard $K{=}20$	—	0.064	0.064	0.158	— / 33 s
synergistic $K{=}5$	0.018	0.018	0.021	—	0.17 s / 1.1 s

Irreducibility vs. optimality. A scripted adversarial instance instantiates the irreducibility-vs-optimality gap: valid deletion sets $\{a, b, c\} \supset \{a, b\} \supset \{c\}$ coexist, the restoration scan (saliency order $a > b > c$) returns the irreducible $\{a, b\}$ (cost 2), and $\{c\}$ (cost 1) is the true minimum—the certificate characterizes the returned set, and the gap is measured exactly where enumeration is feasible (Appendix I).

Synergistic variant. When the mechanism is not learnable (test accuracy $0.25 \approx$ chance; the failure decomposes as compounded per-part errors), all hits collapse to chance on the data-causal target while model-dependence sits at 0.20; the double hit-rate evaluation makes the cause visible—the drop reflects model unlearnability, not explainer error—which a single hit rate conflates.

Scaling. Figure 8 reports the hit rates and wall-clock cost of all search procedures as the relation vocabulary grows; exact search is the fastest search in the benchmark regime ($K{\leq}10$) and certifies the minimum-cost answer at every feasible K, while at $K{=}20$ reliance drifts onto noise relations (test accuracy 0.33) and the budgeted branch-and-bound is the only approximation that improves over greedy/beam (companion implementation).

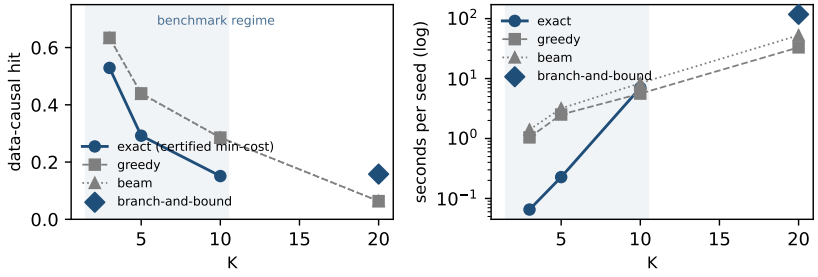


Figure 8: Relation search as the vocabulary grows (synthetic SCM, 10 data seeds, mean $\pm$ 1 std; shaded: benchmark regime $K{\leq}10$). (a) Data-causal hit rate: exact search certifies the minimum-cost answer at every feasible K; the effect-ordered greedy/beam heuristics rank the causal relation higher because the lexicographic cost prefers cheaper noise relations; at $K{=}20$ the budgeted branch-and-bound (bnb) is the only approximation that improves over greedy/beam. (b) Wall-clock per seed: exact search is the fastest in the benchmark regime.