ROBUSTNESS BY CANCELLATION: BREAK, REPAIR, AND SIGNED STRUCTURAL RESPONSES IN GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A graph neural network can appear structurally robust even when many predictions change correctness. Topology perturbations can Break initially correct predictions and Repair initially incorrect ones, while aggregate accuracy records only their net difference. We show that both transitions follow a common margin-crossing account: clean margin sets the crossing threshold, structural response provides the signed displacement, and clean-state opportunity weights conditional susceptibility into population-level flow. On the development graphs, an independently measured response signal improves held-out Break/Repair susceptibility AUC beyond clean margin by 0.058–0.113. On external graphs, the boundary-distance relation remains stable under perturbation-bank resampling, whereas response predictiveness is setting-dependent, including a Repair case at chance. Across competence-filtered datasets and targeted changes in architecture, perturbation operator, and graph scale, Break and Repair recur despite variation in magnitude and net direction. Aggregate accuracy therefore does not, by itself, establish prediction stability.

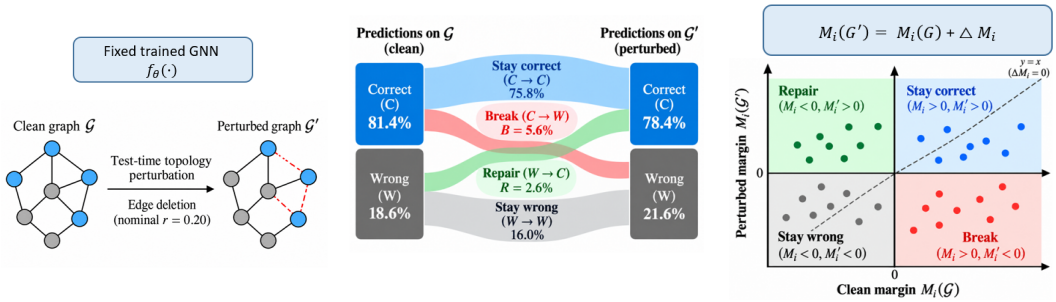
1 INTRODUCTION

A graph neural network can be robust in accuracy while individual predictions remain unstable. Message-passing GNNs couple node representations to graph topology, making predictions sensitive to structural perturbations Kipf & Welling (2016); Hamilton et al. (2017); Veličković et al. (2017); Zügner et al. (2018); Wang et al. (2021). Aggregate accuracy records only the net effect of the resulting correctness changes. In a representative Cora/GCN setting, accuracy drops from 81.4% to 78.4%, while 5.6% of predictions move from correct to wrong and 2.6% from wrong to correct. Thus, 8.2% of predictions change correctness state despite a net decrease of only 3.0 percentage points.

We call the $C \to W$ transition a *Break* and the $W \to C$ transition a *Repair*. Prior work on model updates and prediction churn shows that similar aggregate performance can coexist with substantial sample-level disagreement Yan et al. (2021); Roth & Liebig (2024). We instead keep the trained predictor fixed and perturb only graph topology at test time (Fig. 1a). This setting asks what determines whether the same perturbation process Breaks one prediction while Repairing another. Because aggregate accuracy retains only the net difference between these opposing flows, distinct correctness responses can produce similar aggregate changes (Fig. 1b).

Break and Repair are opposite crossings of the zero true-class margin. Structural edits can induce heterogeneous prediction-level effects Song et al. (2023); Lucic et al. (2022); Zhang et al. (2026); we study how these effects translate into correctness transitions. Let M_i denote the clean true-class margin and ΔM_i its topology-induced displacement. A Break occurs when a positive margin crosses below zero, whereas a Repair occurs when a negative margin crosses above zero (Fig. 1c). Clean-state opportunity determines how these conditional transitions contribute at the population level, giving $\Delta\mathrm{Acc} = (1 - A)R_{\mathrm{cond}} - AB_{\mathrm{cond}}$. This decomposition motivates a stronger empirical test: if structural response contains information beyond clean boundary proximity, then a response measured on a disjoint perturbation bank should improve susceptibility prediction after the clean

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.



(a) Aggregate accuracy hides correctness turnover.

(b) The same perturbation process Breaks and Repairs predictions.

(c) Break and Repair are opposite margin crossings.

Figure 1: **Aggregate accuracy is the net projection of a richer bidirectional correctness response.** (a) We evaluate a fixed trained GNN under a test-time topology perturbation, illustrated here by nominal edge deletion at $r = 0.20$. (b) The same perturbation process simultaneously Breaks initially correct predictions and Repairs initially incorrect predictions. In the canonical Cora/GCN example, $B = 5.6\%$ and $R = 2.6\%$, yielding $\Delta\mathrm{Acc} = R - B = -3.0$ pp despite 8.2% total correctness turnover. (c) Break and Repair correspond exactly to opposite crossings of the zero true-class margin boundary under the signed displacement $\Delta M_i = M_i(G') - M_i(G)$.

margin is known. On Cora and Roman-empire, this response improves Break/Repair AUC by 0.058–0.113. This held-out predictiveness is not implied by the decomposition itself.

External tests separate stable boundary geometry from setting-dependent response persistence. Across PubMed, Minesweeper, and ogbn-arxiv, the boundary-distance relation persists under perturbation-bank resampling, whereas response predictiveness varies by setting, including one Repair result that is unstable around chance. The same account yields an exact class-space distinction: destabilizing an incorrect prediction is equivalent to Repair in binary classification but not in general multiclass prediction. Across competence-filtered datasets and targeted changes in architecture, perturbation operator, and graph scale, both Break and Repair recur, while their magnitudes, balance, and net direction vary; architecture and operator changes can also reverse the sign of $\Delta\mathrm{Acc}$. We therefore claim recurrence of the bidirectional response form within the tested competent regimes, not a fixed robustness level or a universal property. Our contributions are threefold:

- **We reveal robustness by cancellation.** Aggregate accuracy retains only the net balance between Break and Repair and can therefore conceal substantial correctness turnover.
- **We develop a micro-to-macro account of this hidden response.** Clean margin determines boundary position, signed structural response supplies the perturbation-induced displacement, and clean-state opportunity weights conditional susceptibility to produce the aggregate accuracy change.
- **We derive testable consequences and identify the scope of the response structure.** External tests separate stable boundary geometry from setting-dependent response persistence and establish an exact binary–multiclass distinction, while breadth and targeted stress tests show that response form recurs even when magnitude and net direction do not.

2 Related Work

2.1 Structural Robustness under Topology Perturbations

Prior work studies how topology perturbations affect GNN predictions through structural attacks, robustness evaluation, and defenses Zügner et al. (2018); Zhu et al. (2019); Jin et al. (2020); Zhang & Zitnik (2020); Wang et al. (2021). Structural interventions can also improve predictive performance through edge removal or graph transformation during training or at test time Rong et al. (2019); Chen et al. (2020); Jin et al. (2022); Liu et al. (2025); Yu et al. (2026). These studies primarily characterize whether predictive performance degrades, improves, or remains stable. We instead

study the prediction-level correctness process underlying the aggregate change, focusing on the joint Break/Repair response to a common test-time structural intervention.

2.2 Edge Influence and Counterfactual Structural Explanations

Structural influence methods quantify how graph elements affect predictions or use these effects for graph rectification Song et al. (2023); Heo et al. (2026). Counterfactual explainers identify structural edits sufficient to change a prediction Lucic et al. (2022); Zhang et al. (2026). These lines of work show that structural effects can be heterogeneous and directional. We instead characterize signed responses across nodes under a common perturbation process and study how they become margin crossings, correctness transitions, and population-level Break/Repair flows.

2.3 Prediction Flips, Churn, and Correctness Transitions

Correct-to-wrong and wrong-to-correct transitions are established in model-update and prediction-churn settings Yan et al. (2021); Roth & Liebig (2024); Ishimoto et al. (2025). These studies show that similar aggregate performance can conceal substantial sample-level disagreement. We do not claim novelty for the transitions themselves. Our setting keeps the trained GNN fixed and perturbs only graph topology at test time, comparing $f_\theta(G, X)$ with $f_\theta(G', X)$. We connect topology-induced signed margin responses to boundary crossing, Break/Repair susceptibility, and the resulting aggregate accuracy change.

3 Structural Correctness Response

3.1 Problem Setting

We study how a fixed trained GNN responds to a test-time change in graph topology. Let $G = (V, E)$ denote the observed graph with node features X and labels y, and let f_θ be a trained node classifier with prediction $\hat{y}_i(G) = \arg\max_c f_\theta(G, X)_{i,c}$. A structural perturbation transforms G into G' while θ, X, and y remain fixed. Any prediction change therefore reflects the response of the same predictor to altered relational structure.

Let $c_i(G) = \mathbb{I}[\hat{y}_i(G) = y_i]$ denote the correctness state of node i. Each node follows one of four transitions: $C \to C$, $C \to W$, $W \to C$, or $W \to W$. Conventional evaluation compresses these transitions into the scalar change $\Delta\mathrm{Acc} = \mathrm{Acc}(G') - \mathrm{Acc}(G)$. This quantity preserves the net change in correctness but not how correctness is redistributed across predictions.

3.2 Bidirectional Correctness Response

We call the two off-diagonal transitions ***Break*** and ***Repair***,

$$B = P(C \to W), \qquad R = P(W \to C). \tag{1}$$

Break and Repair can arise under the same fixed predictor and perturbation process, as illustrated in Fig. 1b. They therefore represent opposing node-level responses to the same structural intervention rather than distinct perturbation classes. Their balance determines the aggregate change in accuracy.

3.3 Robustness by Cancellation

Break and Repair enter aggregate accuracy with opposite signs:

$$\Delta\mathrm{Acc} = R - B. \tag{2}$$

This identity makes the information loss explicit. The mapping

$$(B, R) \longmapsto R - B$$

is many-to-one, so different correctness flows can produce the same aggregate accuracy change. We use

$$\mathrm{CRM} = B + R \tag{3}$$

only as a bookkeeping quantity for total correctness turnover.

The canonical Cora–GCN example illustrates this cancellation. Accuracy decreases from 81.4% to 78.4%, giving $\Delta\mathrm{Acc} = -3.0$ percentage points, while 5.6% of predictions Break and 2.6% Repair. Thus, 8.2% of predictions change correctness state despite a net decrease of only 3.0 points. The ambiguity also appears when aggregate outcomes are nearly matched. On Amazon-ratings, GCN and GraphSAGE differ by only 0.07 percentage points in $\Delta\mathrm{Acc}$ (-0.28 vs. -0.35) but by 1.65 points in total correctness turnover (3.98% vs. 5.63%).

Aggregate accuracy therefore preserves the imbalance $R-B$ while discarding the amount of correctness movement in either direction. This decomposition explains what aggregate robustness hides, but not why an individual prediction Breaks or Repairs. Section 4 addresses that question through signed changes in the true-class margin.

3.4 Experimental Protocol

We use Cora and Roman-empire for development, and PubMed, Minesweeper, and ogbn-arxiv for external validation. The competence-aware breadth evaluation additionally covers CiteSeer, Amazon-Photo, Amazon-ratings, and Tolokers. Unless stated otherwise, breadth experiments use test-time edge deletion at nominal $r = 0.20$ with model parameters fixed.

GCN and GraphSAGE use two message-passing layers and validation-selected checkpoints after 200 training epochs. Breadth and stress tests use three training seeds, whereas external tests fix the seed-0 checkpoint. For held-out susceptibility analysis, Bank A measures structural response and disjoint Bank B defines Break or Repair susceptibility. The ogbn-arxiv external analysis uses a fixed label-agnostic sample of 5,000 test nodes while retaining the full symmetrized graph for message passing. Appendix A provides complete splits, hyperparameters, perturbation protocols, bank construction, and implementation details.

4 Break and Repair as Margin-Crossing Events

4.1 State and Response in Margin Space

Break and Repair share the same geometry in true-class margin space. For node i, we define the clean true-class margin as

$$M_i(G) = z_{i,y_i}(G) - \max_{c \neq y_i} z_{i,c}(G). \tag{4}$$

Ignoring exact ties, $M_i(G) > 0$ indicates a correct prediction and $M_i(G) < 0$ an incorrect one.[1] A topology perturbation induces the signed margin displacement

$$\Delta M_i = M_i(G') - M_i(G). \tag{5}$$

The clean margin specifies the initial side and distance from the correctness boundary, while ΔM_i specifies the perturbation-induced motion. A Break or Repair occurs exactly when this motion crosses zero:

$$\begin{aligned} \text{Break} &\iff M_i(G) > 0 \ \wedge \ \Delta M_i < -M_i(G), \\ \text{Repair} &\iff M_i(G) < 0 \ \wedge \ \Delta M_i > -M_i(G). \end{aligned} \tag{6}$$

The magnitude $|M_i(G)|$ sets the crossing threshold, while the sign and magnitude of ΔM_i determine the realized structural response (Appendix Fig. 4). Clean margin therefore determines how much directed displacement is required for a transition, but not the displacement itself. Nodes at similar boundary distances can receive different responses, and a node farther from zero can still cross under a sufficiently large displacement.

4.2 Boundary Geometry and Transition Susceptibility

The crossing conditions imply that nodes closer to the zero-margin boundary require less appropriately directed displacement to change correctness. We quantify this threshold by the clean boundary distance

$$d_i = |M_i(G)|. \tag{7}$$

[1] Exact ties satisfy $M_i(G) = 0$ and follow the model's deterministic tie-breaking rule.

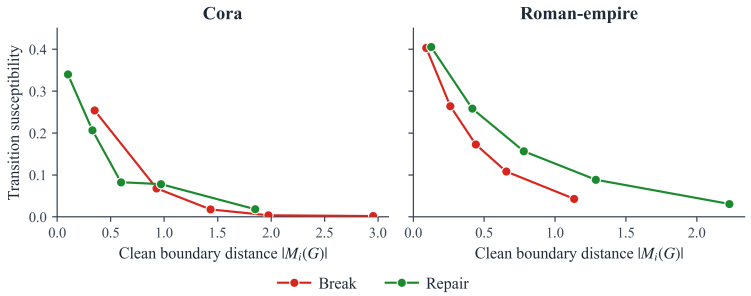


Figure 2: **Crossings concentrate near the boundary.** Break and Repair susceptibility decrease with clean boundary distance on Cora and Roman-empire. Points show seed-averaged values over five equal-count bins; lines are visual guides.

Because only initially correct nodes can Break and only initially incorrect nodes can Repair, we evaluate each transition within its eligible population.

For a clean-distance bin $\mathcal{Q}_k$, we define the conditional susceptibilities

$$\begin{aligned} S_B(\mathcal{Q}_k) &= P(\text{Break} \mid M_i(G) > 0,\, d_i \in \mathcal{Q}_k)\,, \\ S_R(\mathcal{Q}_k) &= P(\text{Repair} \mid M_i(G) < 0,\, d_i \in \mathcal{Q}_k)\,. \end{aligned} \tag{8}$$

We estimate both quantities by empirical transition frequency within the corresponding eligible population.

The development graphs follow the predicted ordering. Break susceptibility decreases with d_i among initially correct nodes, and Repair follows the same pattern among initially incorrect nodes (Fig. 2). The per-node Spearman associations are negative for both transitions across all evaluated seeds. Appendix B.1 reports the per-seed associations and binning details.

Boundary distance therefore characterizes the crossing threshold, but not the sign or magnitude of ΔM_i. Nodes with similar clean boundary geometry can still exhibit different structural responses. The next section examines this topology-induced signed response.

5 Signed Structural Influence

Section 4 specifies the crossing threshold but not the topology-induced displacement ΔM_i. We therefore test whether local structural influence predicts finite-edit direction and whether independently measured structural response adds information beyond clean boundary geometry.

5.1 Local Influence Predicts Finite-Edit Direction

Let A denote the clean adjacency representation and a_e the coordinate of structural element e. We define the signed influence of e on node i as

$$I_{i,e} = \frac{\partial M_i}{\partial a_e}. \tag{9}$$

For a realized edit Δa_e, the product $I_{i,e}\Delta a_e$ predicts the local direction of margin motion. For a general perturbation ΔA, the first-order approximation becomes

$$\Delta M_i \approx \langle \nabla_A M_i, \Delta A \rangle = \sum_e I_{i,e}\Delta a_e. \tag{10}$$

Positive and negative values predict increases and decreases in the true-class margin, respectively. This approximation is local to the clean graph and does not recover the exact response to a finite binary edit.

We evaluate its directional validity by comparing the sign of $I_{i,e}\Delta a_e$ with the exact leave-one-edge-out margin change after recomputing the model. Figure 3 shows sign agreement of 88.0% on Cora and 73.6% on Roman-empire. Local influence therefore remains informative about finite-edit direction, although it does not recover the exact response magnitude. Appendix B.2 gives the sampling and evaluation details.

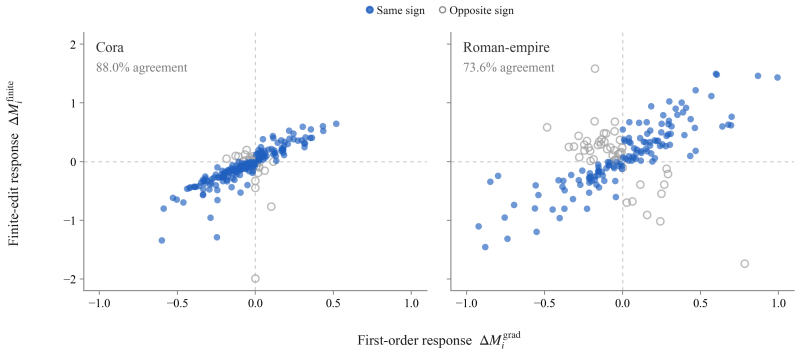


Figure 3: **Local structural influence remains directionally informative under finite edits.** For each evaluated node–edge pair, we compare the sign of the first-order response $I_{i,e}\Delta a_e$ with that of the exact finite leave-one-edge-out margin change after recomputing the model. Filled points denote sign agreement and open points denote disagreement. Agreement reaches 88.0% on Cora and 73.6% on Roman-empire. The comparison concerns response direction rather than exact finite-response magnitude.

Table 1: **Independent structural response adds predictive information beyond clean boundary geometry.** Bank A measures structural response, while disjoint Bank B defines held-out Break or Repair susceptibility.

Dataset	Target	Clean margin	+ Bank-A response	ΔAUC
Cora	Break	0.873	**0.932**	**+0.058**
Cora	Repair	0.821	**0.893**	**+0.072**
Roman-empire	Break	0.783	**0.896**	**+0.113**
Roman-empire	Repair	0.842	**0.903**	**+0.061**

5.2 Independent Response Adds Beyond Boundary Geometry

To separate structural response from clean-margin fragility, we use disjoint perturbation banks: Bank A measures signed mean margin displacement,

$$s_i^{(A)} = \frac{1}{|\mathcal{R}_A|} \sum_{(r,t)\in\mathcal{R}_A} \left[M_i(G'_{r,t}) - M_i(G)\right], \tag{11}$$

while Bank B defines held-out Break and Repair susceptibility among initially correct and incorrect nodes, respectively. The banks share no realization.

Adding Bank-A response to clean margin improves held-out AUC by 0.058–0.113 in all four dataset–transition settings (Table 1). The gain is positive for every training seed and remains positive with degree in the baseline; Appendix B.3 gives per-seed results. Bank-A response is only weakly correlated with clean margin ($|r| \leq 0.19$). Thus, structural response provides predictive information beyond clean boundary geometry. This evidence is node-level and within the same fixed graph and trained model; it does not establish per-event prediction, cross-dataset transfer, or optimality over other perturbation-derived instability summaries.

6 Opportunity–Susceptibility Decomposition

Break and Repair have different eligible populations, so node-level susceptibility alone does not determine aggregate accuracy. We therefore separate conditional susceptibility from clean-state opportunity.

Table 2: **External tests of boundary geometry, response persistence, and Repair directionality.** P1/P2 report mean±SD over three perturbation-bank repeats at a fixed checkpoint (25 Bank-A, 50 Bank-B trials), quantifying bank rather than training-seed variability. P3 compares wrong-state destabilization (D) with Repair (R); full protocols are in Appendix C.

	P1: Boundary		P2: Response		P3: Directionality		
Dataset	ρ_B	ρ_R	$\mathbf{AUC}_B$	$\mathbf{AUC}_R$	$\mathbf{AUC}_D$	$\mathbf{AUC}_R$	$\mathbf{\Delta}_{D-R}$
PubMed	$\mathbf{-0.580 \pm 0.003}$	$\mathbf{-0.575 \pm 0.017}$	$\mathbf{0.709 \pm 0.005}$	$\mathbf{0.768 \pm 0.011}$	0.960	0.926	$\mathbf{+0.033}$
Minesweeper	$\mathbf{-0.543 \pm 0.006}$	$\mathbf{-0.875 \pm 0.006}$	0.526 ± 0.008	$0.498 \pm 0.025^{\dagger}$	0.970	0.970	$\mathbf{0.000}$
ogbn-arxiv	$\mathbf{-0.681 \pm 0.003}$	$\mathbf{-0.665 \pm 0.001}$	$\mathbf{0.662 \pm 0.003}$	$\mathbf{0.637 \pm 0.006}$	0.979	0.729	$\mathbf{+0.249}$

† Minesweeper Repair varies around chance across perturbation banks (range 0.476–0.533). P2 is evaluated within each graph and does not test cross-dataset predictor transfer. N_C/N_W is $794/206$ for PubMed (GCN), $2128/372$ for Minesweeper (GraphSAGE), and $3067/1933$ for ogbn-arxiv (GCN); ogbn-arxiv uses the fixed 5,000-node subset.

6.1 From Susceptibility to Population Flow

Let

$$A = P(M > 0), \qquad 1 - A = P(M < 0), \tag{12}$$

denote the clean-correct and clean-wrong population masses, and define

$$B_{\text{cond}} = P(\Delta M < -M \mid M > 0), \qquad R_{\text{cond}} = P(\Delta M > -M \mid M < 0). \tag{13}$$

The population-level transition rates factor exactly as

$$B = AB_{\text{cond}}, \qquad R = (1 - A)R_{\text{cond}}, \tag{14}$$

which gives

$$\Delta\text{Acc} = (1 - A)R_{\text{cond}} - AB_{\text{cond}}. \tag{15}$$

Aggregate accuracy therefore depends on conditional susceptibility weighted by the available clean-state opportunity.

6.2 The Cancellation Boundary

For fixed $(B_{\text{cond}}, R_{\text{cond}})$, setting Eq. 15 to zero gives

$$A^{\star} = \frac{R_{\text{cond}}}{B_{\text{cond}} + R_{\text{cond}}}, \tag{16}$$

for $B_{\text{cond}} + R_{\text{cond}} > 0$. The cases $A < A^{\star}$, $A = A^{\star}$, and $A > A^{\star}$ yield $\Delta\text{Acc} > 0$, $\Delta\text{Acc} = 0$, and $\Delta\text{Acc} < 0$, respectively. Equivalently,

$$\Delta\text{Acc} = -\left(B_{\text{cond}} + R_{\text{cond}}\right)\left(A - A^{\star}\right). \tag{17}$$

The conditional susceptibilities determine $A^{\star}$, while A specifies the operating point relative to this boundary. The boundary is the exact zero-crossing of the decomposition, not a learned or fitted quantity.

7 Falsifiable Predictions and External Validation

Sections 4–6 yield three testable consequences beyond Cora/Roman-empire. **P1: Boundary geometry.** Within each eligible population, larger clean boundary distance should correspond to lower susceptibility. **P2: Response persistence.** Structural response measured under one perturbation bank can remain informative under a disjoint bank, testing unseen perturbations within the same graph rather than cross-dataset transfer. **P3: Repair directionality.** Destabilization is equivalent to Repair in binary classification, whereas multiclass prediction permits wrong-to-wrong transitions. We test P1–P3 on PubMed, Minesweeper, and ogbn-arxiv.

7.1 Geometry Persists More Consistently Than Response

Table 2 shows that P1 preserves the predicted negative association in all six dataset–transition settings, with perturbation-bank standard deviations no larger than 0.017. P2 is less consistent: the response probe remains informative on PubMed and ogbn-arxiv, whereas Minesweeper is weak for Break (0.526 ± 0.008) and unstable around chance for Repair (0.498 ± 0.025, range 0.476–0.533). Boundary geometry therefore persists more consistently than structural response across the tested external settings. All tests hold the trained checkpoint fixed, so these statistics quantify perturbation-bank rather than training-seed variability.

7.2 Repair Has an Exact Binary–Multiclass Boundary

For an initially incorrect node with clean prediction $a = \hat{y}_i(G) \neq y_i$, define

$$\mathcal{D}_i = \{\hat{y}_i(G') \neq a\}, \qquad \mathcal{R}_i = \{\hat{y}_i(G') = y_i\}. \tag{18}$$

Every Repair destabilizes the original wrong prediction, so $\mathcal{R}_i \subseteq \mathcal{D}_i$. In multiclass prediction, $\mathcal{D}_i \setminus \mathcal{R}_i$ can contain wrong-to-wrong transitions, making destabilization necessary but not sufficient for Repair. When $K = 2$, the true class is the only alternative to a, yielding

$$K = 2 \quad \Longrightarrow \quad \mathcal{D}_i = \mathcal{R}_i \tag{19}$$

for every initially wrong node and perturbation trial.

The P3 results in Table 2 follow this distinction. Minesweeper is binary and yields identical destabilization and Repair AUCs of 0.970, whereas PubMed and ogbn-arxiv are multiclass and yield gaps of $+0.033$ and $+0.249$, respectively. These gaps show that destabilization can include transitions that do not recover correctness in the tested settings, but they do not imply a monotonic relation with class count.

8 Generalization and Scope

Section 7 tests specific consequences of the response account outside the development pair. We now ask whether the bidirectional response form itself recurs across competent predictors and targeted changes in evaluation design. The breadth claim concerns form rather than magnitude: we test whether both Break and Repair recur, not whether their rates, relative balance, or aggregate direction remain fixed.

8.1 Competence-Aware Breadth

We apply the frozen competence audit before interpreting breadth results. Table 3 reports the canonical evaluated setting for each dataset, while Appendix D.1 gives the full candidate-model audit. All seven interpretable settings exhibit both Break and Repair, with CRM ranging from 2.06% to 12.58% and Break–Repair balance varying across datasets. The breadth result is therefore the recurrence of opposing correctness transitions under competent or non-degenerate predictors, not a common robustness level or fixed Break–Repair ratio. Tolokers remains outside this claim because no tested model passes the frozen competence procedure.

8.2 Form Persists, Magnitude Does Not

Table 4 summarizes targeted changes in architecture, perturbation operator, and graph scale; Appendix D.2 reports all 24 canonical cells. Every cell exhibits both Break and Repair, while response magnitude and aggregate direction vary. CRM ranges from 7.7% to 19.3% across architectures, 7.1% to 19.3% across matched-budget operators, and 3.9% to 8.5% on ogbn-arxiv. Architecture and operator changes can also reverse the sign of $\Delta\mathrm{Acc}$. The recurring property is therefore the bidirectional response form, not a fixed magnitude, Break–Repair balance, or net direction.

This recurrence remains limited to the tested regimes. The stress tests are targeted rather than fully crossed, and the synthetic audit includes regimes with very small Repair rates, including individual replicates with no observed Repair (Appendix D.3). We therefore interpret bidirectionality as a recurring response structure across the tested competent and non-degenerate settings, not as a universal property.

Table 3: **Competence-aware breadth of bidirectional structural response.** Results use edge deletion at nominal $r = 0.20$ with three training seeds; Tolokers is unresolved and excluded from positive breadth evidence.

Dataset	Model	Competence	Acc	B	R	CRM	ΔAcc
Cora	GCN	**Competent**	81.57	**5.63**	**2.67**	8.30	-2.96
CiteSeer	GraphSAGE	**Competent**	70.57	**8.01**	**2.59**	10.60	-5.42
PubMed	GCN	**Competent**	78.80	**3.80**	**2.72**	6.52	-1.08
Amazon-Photo†	GraphSAGE	**Competent**	85.98	**1.40**	**0.66**	2.06	-0.75
Roman-empire	GraphSAGE	**Competent**	71.13	**9.24**	**3.34**	12.58	-5.90
Amazon-ratings	GAT64	**Borderline**	43.44	**3.32**	**2.74**	6.06	-0.58
Minesweeper	GraphSAGE	**Borderline**	85.27	**7.20**	**3.00**	10.20	-4.20
Tolokers	N/A	**Unresolved**	N/A	N/A	N/A	N/A	N/A

Accuracy, B, R, and CRM are percentages; ΔAcc is in percentage points. †Amazon-Photo uses the fixed breadth split because the corresponding PyG dataset provides no official split.

Table 4: **Targeted stress tests across architecture, operator, and scale.** Ranges summarize all canonical cells within each stress axis.

Axis	Coverage	B	R	CRM	ΔAcc	Bidirectional
Architecture	4 arch. × 2 graphs	5.6–9.2	2.0–11.4	7.7–19.3	-5.8–$+3.5$	**8/8**
Operator	3 op. × 2 models × 2 graphs	5.2–9.8	1.9–10.8	7.1–19.3	-5.8–$+2.4$	**12/12**
Scale	2 models × 2 rates	2.2–5.7	1.7–2.8	3.9–8.5	-2.9–-0.5	**4/4**

B, R, and CRM are percentages; ΔAcc is in percentage points. Architecture covers GCN, GraphSAGE, GAT, and APPNP on Cora/Roman-empire; operators use matched edit budgets ($q_{\text{edit}} \approx 0.40$) and test persistence, not ranking; ogbn-arxiv is a large-graph stress test, not a scale sweep.

9 Discussion and Limitations

Aggregate robustness does not imply prediction stability because accuracy records only the net balance between Break and Repair. Our results separate this projection into clean boundary geometry, topology-induced structural response, and clean-state opportunity, while showing that bidirectionality recurs more consistently than its magnitude, Break–Repair balance, or aggregate direction. Our conclusions are limited to fixed trained models under test-time topology perturbations for node classification. The architecture, operator, and dataset coverage is targeted rather than exhaustive; signed structural influence is only a local first-order approximation to finite edits; perturbation-bank analyses hold each checkpoint fixed; and held-out probes test within-graph persistence rather than cross-dataset transfer. Tolokers remains unresolved, GAT and APPNP are evaluated only on Cora and Roman-empire, ogbn-arxiv provides a large-graph stress test rather than a scale sweep, and the synthetic audit includes regimes with very small Repair rates, including individual replicates with no observed Repair. We therefore interpret bidirectionality as a recurring response form within the tested regimes, not as a universal property.

10 Conclusion

Aggregate structural robustness can conceal substantial bidirectional prediction-level change. Test-time topology perturbations can both Break initially correct predictions and Repair initially incorrect ones, while aggregate accuracy records only their net balance. These transitions follow a common margin-crossing account: clean margin sets the crossing threshold, structural response determines margin motion, and clean-state opportunity weights conditional susceptibility into population-level flows. Across the tested competent regimes, bidirectionality recurs even as its magnitude, Break–Repair balance, and aggregate direction vary. Aggregate accuracy therefore does not, by itself, establish prediction stability; a small net change can also reflect cancellation between perturbation-induced failures and recoveries.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and to support the retrieval and discovery of related work. All AI-assisted text was reviewed and revised by the authors. Literature identified through AI-assisted retrieval was checked against the original sources before being cited or used in the manuscript. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the experimental details needed to reproduce the reported results in the appendices. Appendix A specifies the datasets and splits, model and training configurations, structural perturbation protocols, perturbation-bank construction, competence criteria, software versions, hardware, and random-seed handling. Appendix B provides the protocols for the development-graph analyses, Appendix C details the external-validation procedure, and Appendix D reports the competence audit, stress tests, and synthetic scope analysis. We explicitly distinguish training-seed variability from perturbation-bank variability and document the determinism limitations of the original GPU implementation together with the subsequent CPU audit recomputations.

REFERENCES

Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems*, 33:19314–19326, 2020.

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

Jaeseung Heo, Kyeongheung Yun, Seokwon Yoon, MoonJeong Park, Jungseul Ok, and Dongwoo Kim. Influence functions for edge edits in non-convex graph neural networks. *Advances in Neural Information Processing Systems*, 38:166208–166234, 2026.

Yuta Ishimoto, Masanari Kondo, Lei Ma, Naoyasu Ubayashi, and Yasutaka Kamei. Repairs and breaks prediction for deep neural networks. *ACM Transactions on Software Engineering and Methodology*, 34(4):1–42, 2025.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 66–74, 2020.

Wei Jin, Tong Zhao, Jiayuan Ding, Yozen Liu, Jiliang Tang, and Neil Shah. Empowering graph representation learning with test-time graph transformation. *arXiv preprint arXiv:2210.03561*, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Xin Liu, Yuxiang Zhang, Meng Wu, Mingyu Yan, Kun He, Wei Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Revisiting edge perturbation for graph neural network in graph data augmentation and attack. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):4225–4238, 2025.

Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *International conference on artificial intelligence and statistics*, pp. 4499–4511. PMLR, 2022.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Andreas Roth and Thomas Liebig. Distilling influences to mitigate prediction churn in graph neural networks. In *Asian Conference on Machine Learning*, pp. 1151–1166. PMLR, 2024.

Jaeyun Song, SungYub Kim, and Eunho Yang. Rge: A repulsive graph rectification for node classification via influence. In *International Conference on Machine Learning*, pp. 32331–32348. PMLR, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1645–1653, 2021.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14299–14308, 2021.

Dongyang Yu, Xia Cui, and Rong Xiao. Setta: Parameter-free test-time adaptation for graph neural networks via spectral-energy-guided semantic refinement. *Big Data and Cognitive Computing*, 10(8):260, 2026.

Xiang Zhang and Marinka Zitnik. Gnnguard: Defending graph neural networks against adversarial attacks. *Advances in neural information processing systems*, 33:9263–9275, 2020.

Yu Zhang, Bin Yang, Arijit Khan, and Cuneyt Akcora. Atex-cf: attack-informed counterfactual explanations for graph neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 118042–118055, 2026.

Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1399–1407, 2019.

Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2847–2856, 2018.

A Experimental Setup

A.1 Datasets and Splits

We evaluate nine node-classification datasets. Cora and Roman-empire are used for development; PubMed, Minesweeper, and ogbn-arxiv for external validation; and CiteSeer, Amazon-Photo, Amazon-ratings, and Tolokers extend the competence-aware breadth evaluation. Table 5 summarizes the dataset statistics and evaluation splits.

Cora, CiteSeer, and PubMed use the public Planetoid splits. Roman-empire, Amazon-ratings, Minesweeper, and Tolokers use official split index 0. Amazon-Photo uses a fixed per-class stratified 60/20/20 split with seed 13000. For ogbn-arxiv, we symmetrize the graph before perturbation and retain the full graph for message passing. External validation evaluates a fixed label-agnostic sample of 5,000 test nodes drawn without replacement with seed 60000, whereas the breadth and large-graph stress tests use the full test set.

A.2 Models and Training

GCN and GraphSAGE use two message-passing layers with hidden dimension 16, ReLU, and dropout 0.5. We train each model for 200 epochs with Adam, learning rate 0.01, weight decay 5×10^{-4}, and cross-entropy loss, and select checkpoints by validation accuracy. Breadth and stress-test cells use seeds $0, 1, 2$, while external validation uses the fixed seed-0 checkpoint.

The architecture stress test additionally evaluates GAT and APPNP on Cora and Roman-empire. GAT uses four attention heads of width 16, and APPNP uses hidden dimension 16 with $K = 10$ and teleport probability $\alpha = 0.1$. A wider GAT variant with four heads of width 64 is used only for the competence-repair audit on Amazon-ratings, Minesweeper, and Tolokers.

Table 5: **Datasets and evaluation splits.** Edge counts denote canonical undirected pairs.

Dataset	Classes	Nodes	Edges	Features	Train/Val/Test
Cora	7	2,708	5,278	1,433	140/500/1,000
CiteSeer	6	3,327	4,552	3,703	120/500/1,000
PubMed	3	19,717	44,324	500	60/500/1,000
Amazon-Photo	8	7,650	119,081	745	4,591/1,530/1,529
Roman-empire	18	22,662	32,927	300	11,331/5,665/5,666
Amazon-ratings	5	24,492	93,050	300	12,246/6,123/6,123
Minesweeper	2	10,000	39,402	7	5,000/2,500/2,500
Tolokers	2	11,758	519,000	10	5,879/2,939/2,940
ogbn-arxiv	40	169,343	1,157,799	128	90,941/29,799/48,603

A.3 Structural Perturbations and Banks

All perturbations are applied at test time with model parameters, node features, and labels fixed. Operations act on canonical undirected edge pairs and exclude self-loops and duplicate additions. Deletion samples existing pairs without replacement, addition samples new pairs from the graph complement, and rewiring removes and adds the same number of pairs.

The nominal perturbation rate r is defined relative to the directed edge count, whereas edits are applied to undirected pairs. On the symmetric graphs used here, deletion or addition at $r = 0.20$ therefore corresponds to $q_{\text{edit}} \approx 0.40$. The matched operator comparison uses deletion and addition at $r = 0.20$ and rewiring at $r = 0.10$, yielding approximately matched edit budgets.

Held-out susceptibility analyses use separate perturbation banks. Bank A measures the mean signed true-class margin response, while Bank B defines Break or Repair susceptibility. Development experiments use 50 realizations in each bank. Canonical external experiments use 50/50 realizations on PubMed and Minesweeper and 30/30 on ogbn-arxiv. The independently seeded repeats used for the uncertainty estimates in Table 2 use 25 Bank-A and 50 Bank-B realizations at a fixed checkpoint. Appendix C gives the complete external-validation protocol.

A.4 Competence and Reproducibility

We apply the frozen competence audit before interpreting breadth results. A model is competent if its clean accuracy exceeds the majority-class baseline by at least 10 percentage points, it predicts at least $\min(3, K)$ classes with each represented by at least 1% of predictions, and no seed assigns at least 95% of predictions to a single class. A borderline setting fails one of the first two conditions but remains above the majority baseline, predicts at least two classes, and avoids collapse. Five datasets admit a competent canonical setting, Amazon-ratings and Minesweeper remain borderline but non-degenerate, and Tolokers remains unresolved. Appendix D.1 reports the full candidate-model audit and Tolokers repair attempts.

The original training and Rounds 1–4C evaluations used Python 3.11.10, PyTorch 2.4.1+cu124, PyG 2.8.0.post1, and OGB 1.3.6 on one NVIDIA RTX 4090. The training code seeds Python, NumPy, PyTorch, and CUDA RNGs, while perturbation banks use independent seeded generators. The original implementation does not enforce deterministic GPU algorithms; the recorded numerical scatter did not change argmax predictions. Later CPU-only audit recomputations reproduce the frozen Table 1 values.

B Additional Development-Graph Analyses

This section provides additional protocol details and statistics for the Cora and Roman-empire development analyses.

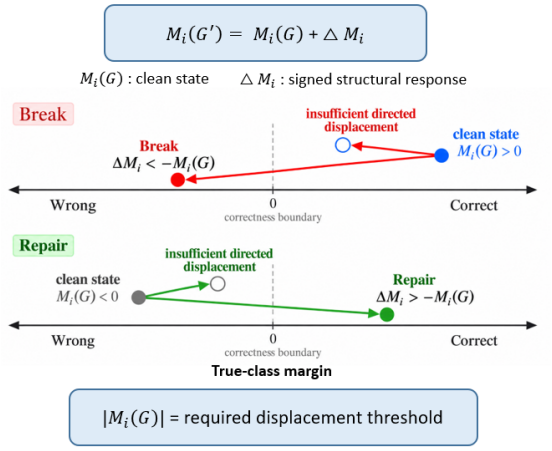


Figure 4: **Clean margin sets the crossing threshold.** For a fixed clean state, an appropriately directed margin displacement changes correctness only when it exceeds the distance to the zero-margin boundary; smaller displacements remain on the original side.

Table 6: **Per-seed boundary-distance associations.** Entries are node-level Spearman correlations.

Dataset	Target	Seed 0	Seed 1	Seed 2
Cora	Break	-0.615	-0.617	-0.639
Cora	Repair	-0.661	-0.559	-0.620
Roman-empire	Break	-0.528	-0.526	-0.555
Roman-empire	Repair	-0.625	-0.619	-0.631

B.1 Boundary-Distance Analysis

For Figure 2, we form five equal-count clean-distance bins separately for initially correct and initially incorrect test nodes within each training seed. Break and Repair susceptibility are the corresponding Bank-B transition frequencies. The plotted curves average the bin distances and transition frequencies across seeds, while the reported Spearman correlations are computed directly at the node level. All twelve seed-level associations are negative. The observed boundary-distance relation therefore does not depend on the visual binning used in Figure 2.

B.2 Finite-Edit Influence Analysis

For Figure 3, we sample 60 test nodes per dataset from the seed-0 GCN checkpoint and evaluate all incident canonical edge pairs, yielding 216 pairs on Cora and 178 on Roman-empire. For each pair, the first-order response is $I_{i,e}\Delta a_e$ with $\Delta a_e = -1$, and the finite response is the exact change in the true-class margin after removing the edge pair and recomputing the model. Sign agreement is $190/216 = 88.0\%$ on Cora and $131/178 = 73.6\%$ on Roman-empire. One exactly zero first-order response on Cora is counted as disagreement. This analysis uses a single checkpoint and evaluates directional agreement rather than uncertainty across trained models.

B.3 Independent Structural-Response Probe

Table 1 uses one fixed disjoint Bank-A/Bank-B pair for each development graph, with 50 deletion realizations per bank. Bank A defines the mean signed true-class margin response, while Bank B defines whether an eligible node has nonzero Break or Repair susceptibility. This binary target is defined relative to the fixed 50-trial Bank-B protocol and should not be interpreted as an estimate of the underlying transition probability. We fit the probe separately for each dataset and transition.

The gain is positive for every training seed in all four settings. These statistics measure variation across independently trained checkpoints under one fixed Bank-A/Bank-B pair; they do not mea-

Table 7: **Per-seed gain from independent structural response.** $\Delta\mathrm{AUC}$ is the gain from adding the Bank-A response to clean margin.

Dataset	Target	Seed 0	Seed 1	Seed 2	Mean±SD
Cora	Break	0.0522	0.0693	0.0533	0.0583 ± 0.0096
Cora	Repair	0.0381	0.0899	0.0878	0.0719 ± 0.0293
Roman-empire	Break	0.1169	0.1211	0.1015	0.1132 ± 0.0103
Roman-empire	Repair	0.0602	0.0602	0.0635	0.0613 ± 0.0019

sure perturbation-bank variability. The unrounded seed-averaged gains are 0.058284, 0.071934, 0.113164, and 0.061311, which yield the rounded values in Table 1.

C External Validation Details

C.1 External Settings and Bank Resampling

All external tests use edge deletion at nominal $r = 0.20$ with a fixed seed-0 checkpoint: GCN for PubMed and ogbn-arxiv, and GraphSAGE for Minesweeper. PubMed and Minesweeper use their full test sets, while ogbn-arxiv uses the fixed 5,000-node subset defined in Appendix A.1. The canonical banks contain 50 Bank-A and 50 Bank-B realizations for PubMed and Minesweeper and 30 realizations per bank for ogbn-arxiv.

The P1/P2 uncertainty estimates use three independently seeded Bank-A/Bank-B pairs at the same fixed checkpoint. Each repeat contains 25 Bank-A and 50 Bank-B realizations. The reported standard deviation is computed across the three repeats and therefore measures perturbation-bank variability rather than training-seed variability. These repeat banks use different trial counts from the canonical banks.

C.2 Boundary Geometry and Response Persistence

P1 evaluates boundary geometry directly at the node level. For initially correct nodes, we compute the Spearman correlation between the clean true-class logit margin M_i and Bank-B Break frequency. For initially incorrect nodes, we compute the corresponding correlation between $-M_i$ and Bank-B Repair frequency. No binning is used.

P2 evaluates whether structural response measured from Bank A remains informative about susceptibility under disjoint Bank B. The feature is the mean Bank-A displacement in the true-class logit margin, and the target indicates whether the corresponding Bank-B Break or Repair rate is nonzero. We standardize the feature and evaluate logistic-regression ROC-AUC with five-fold stratified cross-validation. Each probe is fitted separately within one dataset and transition; no predictor is transferred across datasets.

C.3 Repair Directionality

For an initially incorrect node with clean predicted class a and true class y_i, P3 defines wrong-state destabilization as $D_i = \mathbb{I}[\hat{y}_i(G') \neq a]$ and Repair as $R_i = \mathbb{I}[\hat{y}_i(G') = y_i]$. Thus, $R_i \subseteq D_i$, with equality when $K = 2$. Minesweeper is binary, so $D_i = R_i$ for every initially incorrect node and perturbation trial.

P3 uses the canonical Bank-A label-free eight-feature probe rather than the single true-class-margin response used in P2. For each target, we median-split the per-node Bank-B rate and evaluate five-fold logistic-regression ROC-AUC. The reported gap is $\mathrm{AUC}_D - \mathrm{AUC}_R$. P3 uses the canonical banks and is not averaged over the fresh bank repeats.

Table 8: **Competence audit across breadth datasets.**

Dataset	GCN	GraphSAGE
Cora	Competent	Competent
CiteSeer	Competent	Competent
PubMed	Competent	Competent
Amazon-Photo	Borderline	Competent
Roman-empire	Competent	Competent
Amazon-ratings	Borderline	Borderline
Minesweeper	Incompetent	Borderline
Tolokers	Incompetent	Incompetent

D BREADTH, STRESS TESTS, AND SCOPE

D.1 FULL COMPETENCE AUDIT

We apply the frozen competence criteria from Appendix A.4 before interpreting breadth results. Table 8 summarizes the candidate-model audit and the additional GAT64 repair attempts used when no competent GCN or GraphSAGE model is available.

Five datasets admit at least one competent candidate. We additionally evaluate GAT64 for Amazon-ratings, Minesweeper, and Tolokers because no competent GCN or GraphSAGE setting is available. GAT64 remains borderline on Amazon-ratings and does not reach competent status on Minesweeper or Tolokers. Tolokers remains unresolved: the original GCN and GraphSAGE models collapse toward the majority class, the class-weighted GAT64 avoids collapse but falls below the majority baseline, and the unweighted GAT64 collapses again. We therefore exclude Tolokers from positive breadth evidence.

When multiple candidates are competent, Table 3 reports the frozen canonical evaluated setting. We do not introduce a post hoc rule for selecting among competent candidates.

D.2 ARCHITECTURE, OPERATOR, AND SCALE STRESS TESTS

Table 9 expands the 24 canonical cells summarized in Table 4. The architecture axis evaluates GCN, GraphSAGE, GAT, and APPNP on Cora and Roman-empire. The operator axis evaluates deletion, addition, and rewiring for GCN and GraphSAGE on the same graphs, using deletion and addition at nominal $r = 0.20$ and rewiring at $r = 0.10$ to match $q_{\text{edit}} \approx 0.40$. The large-graph stress test evaluates GCN and GraphSAGE on ogbn-arxiv at $r \in \{0.10, 0.20\}$ using the full test set.

Both Break and Repair occur in all $8/8$ architecture cells, all $12/12$ matched-budget operator cells, and all $4/4$ ogbn-arxiv cells. Their magnitudes and net directions vary, and architecture or operator changes can reverse the sign of ΔAcc. The operator results therefore establish persistence under matched edit budgets rather than a ranking among operators. The ogbn-arxiv results provide a large-graph stress test rather than a scale sweep.

The architecture cells combine frozen results from the original architecture and operator evaluations. GAT and APPNP are tested only on Cora and Roman-empire and were not included in the formal competence audit. The operator-axis deletion cells use the preserved WP5 cell-level outputs; the corresponding original delete-bank files were subsequently overwritten.

D.3 SYNTHETIC SCOPE AUDIT

We use a contextual stochastic block model to probe the scope of bidirectionality under controlled graph regimes. Each graph contains $n = 1500$ nodes in $K = 5$ balanced classes with 32-dimensional features $x_i = \rho\mu_{y_i} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, I)$. We vary homophily $h \in \{0.2, 0.4, 0.6, 0.8\}$, mean degree $d \in \{4, 8, 16\}$, and feature signal $\rho \in \{0.5, 1, 2\}$, yielding 36 regimes with three replicates each. Each replicate uses a 20/20/60 split, a two-layer GCN trained for 150 epochs, and 50 edge-deletion trials at nominal $r = 0.20$.

Table 9: **Canonical stress-test cells.** B, R, and CRM are percentages; $\Delta\mathrm{Acc}$ is in percentage points.

Axis	Dataset	Setting	B	R	CRM	ΔAcc
Architecture	Cora	GCN	5.55	2.57	8.12	-2.98
	Cora	GraphSAGE	7.36	2.84	10.20	-4.52
	Cora	GAT	6.08	2.92	9.00	-3.17
	Cora	APPNP	5.63	2.03	7.66	-3.60
	Roman-empire	GCN	8.44	10.83	19.27	+2.38
	Roman-empire	GraphSAGE	9.18	3.34	12.52	-5.84
	Roman-empire	GAT	7.86	11.41	19.27	+3.55
	Roman-empire	APPNP	8.66	8.11	16.76	-0.55
Operator	Cora	GCN, delete	5.55	2.57	8.12	-2.98
	Cora	GCN, add	7.61	3.00	10.61	-4.61
	Cora	GCN, rewire	7.33	2.87	10.20	-4.45
	Cora	GraphSAGE, delete	7.20	2.72	9.92	-4.48
	Cora	GraphSAGE, add	6.16	2.23	8.39	-3.94
	Cora	GraphSAGE, rewire	7.24	2.43	9.67	-4.81
	Roman-empire	GCN, delete	8.44	10.83	19.27	+2.38
	Roman-empire	GCN, add	9.84	7.98	17.82	-1.85
	Roman-empire	GCN, rewire	8.54	9.08	17.63	+0.54
	Roman-empire	GraphSAGE, delete	9.18	3.38	12.57	-5.80
	Roman-empire	GraphSAGE, add	5.19	1.90	7.09	-3.29
	Roman-empire	GraphSAGE, rewire	6.94	2.46	9.39	-4.48
Large graph	ogbn-arxiv	GCN, $r = 0.10$	2.25	1.70	3.95	-0.54
	ogbn-arxiv	GCN, $r = 0.20$	4.19	2.80	6.98	-1.40
	ogbn-arxiv	GraphSAGE, $r = 0.10$	2.88	1.73	4.61	-1.15
	ogbn-arxiv	GraphSAGE, $r = 0.20$	5.72	2.78	8.49	-2.94

Table 10: **Synthetic regimes with very small Repair.** Entries are means over three replicates; all four regimes have clean accuracy of at least 97.5%.

h	d	ρ	B	R
0.8	8	2.0	0.0455	0.0052
0.8	16	0.5	0.1079	0.0040
0.8	16	1.0	0.0289	0.0009
0.8	16	2.0	0.0126	0.0005

Repair becomes very small in several high-homophily regimes. Table 10 reports the four regime means with $R < 0.006$.

Exact $R = 0$ occurs in only $2/108$ replicate-level rows, both at $h = 0.8$, $d = 16$, and replicate 0, with $\rho = 1.0$ or 2.0. No regime-level mean has zero Repair. The synthetic audit therefore bounds the empirical recurrence claim: bidirectionality can become highly asymmetric, and individual realizations can exhibit no observed Repair. We do not treat these synthetic results as positive breadth evidence.