PROTOCOL-CONDITIONAL ATTRIBUTION IN FAIRNESS-AWARE GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness-aware graph neural networks (GNNs) are typically evaluated by comparing a complete method with an external baseline, even when the scientific claim concerns an intervention intended to reduce bias. Such comparisons show what the complete method does, but not how much of the observed change is associated with the claimed intervention. We separate the complete-method contrast from a matched intervention contrast between intervention-disabled and intervention-enabled versions of the same method, and treat this intervention attribution as conditional on the protocol used to train, select, and evaluate the model, an approach we call protocol-conditional attribution. Across 36 controlled cells spanning eleven methods, the surrounding package has a larger absolute contrast than the claimed intervention in 26/36 cells for demographic parity and 29/36 for equal opportunity (about one half against a more strongly trained baseline), and the intervention contrast is resolved in only 8/36 and 9/36 cells, with most remaining fairness contrasts inconclusive rather than negligible. Changes in method configuration and checkpoint selection frequently alter the magnitude and resolution of intervention attribution, although resolved directions are rarely reversed. Targeted analyses show that longer training can change the intervention contrast at later checkpoints and that an explicit fairness correction can contribute little relative to preceding components. Claims about why a system is fairer therefore require matched intervention evidence under an explicit protocol. Our code is available at `https://anonymous.4open.science/r/FairGNN-Eval-ICLR`.

1 INTRODUCTION

Graph Neural Networks (GNNs) are increasingly used in consequential applications such as financial fraud detection, clinical risk prediction, and social-network analysis (Liu et al., 2021; Boll et al., 2024; Saxena et al., 2024). In these settings, graph structure creates a distinctive fairness challenge: message passing can propagate sensitive-group information encoded in the topology into learned representations (Dong et al., 2022b; 2023a), and disparities can persist even when sensitive attributes are removed from the input features because they remain recoverable from graph structure (Agarwal et al., 2021; Laclau et al., 2022). This has motivated a literature on fairness-aware graph learning.

Fairness-aware GNNs address this problem through diverse mechanisms, including adversarial learning (Dai & Wang, 2021), representation regularization (Agarwal et al., 2021), graph modification or augmentation (Dong et al., 2022a; Ling et al., 2023), sensitive-information suppression or neutralization (Wang et al., 2022; Yang et al., 2024), rebalancing (Li et al., 2024), balanced or fairness-aware message passing (Lin et al., 2024; Jiang et al., 2024), and stability-constrained learning (Chen et al., 2025). Despite these differences, each method designates a component, such as a loss term, a data or graph edit, or a training procedure, as responsible for reducing unfairness; we call this component the method's *claimed fairness intervention* and the released method as a whole the *complete method*. Evaluating such a method therefore involves two questions:

- **RQ1:** Is the complete method fairer than a reference system?
- **RQ2:** Does the claimed fairness intervention contribute to that change?

Standard evaluation answers RQ1. A fairness-aware method is trained alongside baselines, such as a vanilla GCN and earlier fairness-aware methods, on benchmark graphs such as German credit,

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.

Bail, and Pokec, and the complete methods are compared on utility (e.g., AUC) and group disparity (e.g., demographic parity and equal opportunity) (Dai & Wang, 2021; Agarwal et al., 2021; Dong et al., 2022a; 2025); recent benchmarks have made such comparisons more reliable by standardizing pipelines, baselines, and model selection (Han et al., 2024; Cruz & Hardt, 2024; Tan et al., 2026). This evidence establishes whether a complete method is fairer, but not why. A complete method typically differs from the baseline not only in its fairness intervention but also in architecture, preprocessing, augmentation, or optimization, so a favorable comparison is compatible with an intervention that contributes substantially, contributes little, or moves the outcome in the opposite direction; likewise, a utility–fairness trade-off observed for the complete method need not be the trade-off induced by the intervention. Attributing an improvement to the proposed intervention therefore requires evidence for RQ2.

Our central premise is that RQ2 is a different evaluation question from RQ1 and therefore requires different evidence. The relevant comparison is not between the complete method and an external baseline, but between the same method with and without its claimed intervention, with everything else held fixed; only this within-method contrast speaks to the mechanism the method proposes. Moreover, this contrast is measured under a particular way of training, selecting, and evaluating models, so an answer to RQ2 is a property of a method under a *specification*, that is, a method configuration together with a training, selection, and evaluation protocol, rather than of the method name. Component ablations use a similar operation, but typically as supporting evidence under the authors' own protocol; we instead treat the within-method contrast as the primary evidence for an intervention-level claim, relate it to the complete-method change, and ask how stable it is across defensible specifications. We call this approach *protocol-conditional attribution*.

We apply this view to eleven fairness-aware GNN methods on 36 method–dataset pairs (*cells*). Unlike the common practice of averaging repeated runs on a single data split (Agarwal et al., 2021), each cell varies both the data split and the initialization, so that the estimates reflect both sources of variation (Bouthillier et al., 2021). We find that complete-method fairness changes are a poor guide to intervention contributions. The change attributable to the *surrounding package*, i.e., the method's components other than the intervention, exceeds that of the intervention in 72% of cells for demographic parity and 81% for equal opportunity (26 and 29 of 36 cells), a proportion that falls to about one half when the external baseline is trained more strongly. The intervention contrast itself is clearly resolved, i.e., distinguishable from zero at a practically relevant size, in only 22% and 25% of cells (8 and 9 of 36); resolved contrasts almost always reduce disparity, and most remaining ones are inconclusive rather than negligible. We also find that whether an intervention appears to matter depends on the specification: alternative method configurations and evaluation protocols frequently change the magnitude and resolution of the intervention contrast, although they rarely reverse a clearly resolved direction, and targeted analyses trace selected shifts to later training stages or to components that precede the fairness-labeled intervention. Our contributions are threefold:

- **(i) An evaluation question and its evidence:** we frame the evaluation of fairness interventions around RQ2 and use the matched within-method contrast, related to the complete-method change through an exact decomposition, as its primary evidence.
- **(ii) Protocol-conditional attribution:** we define this evidence conditional on the evaluation protocol and method configuration, separating stochastic uncertainty from sensitivity to these choices.
- **(iii) An audit of eleven methods:** we show that complete-method and intervention-level evidence frequently diverge and that intervention-level conclusions depend on the evaluated specification.

2 Related Work

2.1 Fairness-Aware Graph Learning

Fairness-aware graph learning mitigates bias by intervening at different stages of the learning pipeline. Some methods modify the data, features, or graph structure (Dong et al., 2022a; 2023b; Loveland et al., 2022); others impose fairness-related constraints during training (Dai & Wang, 2021; Wang et al., 2022; Chen et al., 2025); and a class of approaches uses sampling, rebalancing, alignment, invariance, or related regularization objectives (Lin et al., 2024; Li et al., 2024; Yang

et al., 2024; Ma et al., 2022; Agarwal et al., 2021). FMP (Jiang et al., 2024) provides a complementary architectural example that separates graph propagation from an explicit fairness-oriented correction, while FnRGNN (Park & Lim, 2025) combines distribution-aware fairness components and is evaluated separately in our task-adapted analysis. Although these mechanisms differ in where and how they intervene, each has a claimed fairness intervention in the sense of Section 1. These categories are descriptive rather than mutually exclusive; we use the same grouping only to organize the empirical analysis.

2.2 Fairness Evaluation and Benchmarking

A growing body of work has examined how experimental design affects the reliability of fairness comparisons. Han et al. (2024) standardize evaluation for in-processing group-fairness methods, while Cruz & Hardt (2024) show that comparisons can be confounded by differences in unconstrained baselines or fairness-constraint relaxation. Tan et al. (2026) further demonstrate that hyperparameter tuning and model selection can materially affect both utility and disparity. In graph learning, PyGDebias (Dong et al., 2024) provides a unified implementation framework, and the fairness-aware graph learning benchmark of Dong et al. (2025) compares representative methods under a common evaluation protocol. These efforts strengthen the evidence for RQ1; the quantity we target is the intervention contribution addressed by RQ2.

2.3 Intervention Attribution and Protocol Dependence

Component ablation is widely used to assess model components, and its conclusions depend on how the ablated state is constructed (Li & Janson, 2024); fairness-aware graph methods commonly report such ablations to support proposed modules (Wang et al., 2022; Li et al., 2024; Jiang et al., 2024). Our matched contrast uses the same operation but differs from such ablations in three concrete respects. First, it is paired with the external-baseline contrast, so that the complete-method change is apportioned between the claimed intervention and the surrounding package (Eq. 6) rather than the component being assessed in isolation. Second, it is estimated for every method under one controlled protocol with nested uncertainty, making contributions comparable across methods. Third, it is treated as conditional on method configuration and protocol, and we measure how it moves when either changes.

Prior work shows that model selection can change empirical conclusions (Cawley & Talbot, 2010) and that baselines, hyperparameters, training procedures, and model selection can change fairness conclusions (Cruz & Hardt, 2024; Han et al., 2024; Tan et al., 2026), and multiverse and specification-curve analyses study variation across defensible analytical choices (Steegen et al., 2016; Simonsohn et al., 2020); in our setting, the specifications are training and selection protocols for learned models, and the quantity that varies is the intervention contrast. Run-to-run variance from data sampling, initialization, and hyperparameter choice can also dominate differences between methods (Bouthillier et al., 2021); we therefore resample splits and runs hierarchically and treat contrasts below a magnitude threshold as unresolved.

3 Protocol-Conditional Attribution of Fairness Interventions

3.1 Evaluation Setting and Outcome Coordinates

We consider binary node classification on an attributed graph $G = (V, E, X)$, where each evaluated node has a binary target y_i and binary sensitive attribute z_i. Let M denote a learning method and P a fixed training, model-selection, and evaluation protocol. An experimental realization is indexed by $\omega = (s, r)$, where s denotes a data split and r a stochastic training run nested within that split. All intervention comparisons are matched within the same realization ω. Compared systems therefore share the same data split and run index. We evaluate predictive utility with AUC and group fairness with demographic parity (DP) and equal opportunity (EO) (Hardt et al., 2016; Mehrabi et al., 2021). For a common higher-is-better orientation, define

$$Y_\omega(M; P) = [\mathrm{AUC}_\omega(M; P), -\Delta_{\mathrm{DP},\omega}(M; P), -\Delta_{\mathrm{EO},\omega}(M; P)]\,. \tag{1}$$

The three coordinates are analyzed separately rather than combined into a scalar utility–fairness objective. Exact metric and decision-rule definitions are given in Appendix D.1.

3.2 Matched Intervention Attribution

For a fairness-aware method, let I denote its claimed fairness intervention. We distinguish an external baseline B, a method-specific system M^{-I} in which I is disabled, and the corresponding system M^{+I} in which I is enabled. We refer to M^{+I} as the complete method and to the components of M^{+I} other than I as its surrounding package. The pair M^{-I} and M^{+I} is matched in all scientifically separable aspects not constitutive of I, including architecture, preprocessing, optimization, and evaluation protocol. Thus, M^{-I} is a method-specific intervention-disabled control rather than a generic unconstrained baseline. For a fixed protocol P, define

$$\tau_{B\to -I}(P) = \mathbb{E}_\omega \left[Y_\omega(M^{-I}; P) - Y_\omega(B; P) \right], \tag{2}$$

$$\tau_{-I\to +I}(P) = \mathbb{E}_\omega \left[Y_\omega(M^{+I}; P) - Y_\omega(M^{-I}; P) \right], \tag{3}$$

$$\tau_{B\to +I}(P) = \mathbb{E}_\omega \left[Y_\omega(M^{+I}; P) - Y_\omega(B; P) \right]. \tag{4}$$

By construction,

$$B \xrightarrow{\tau_{B\to -I}(P)} M^{-I} \xrightarrow{\tau_{-I\to +I}(P)} M^{+I}, \qquad B \xrightarrow{\tau_{B\to +I}(P)} M^{+I}, \tag{5}$$

and therefore

$$\tau_{B\to +I}(P) = \tau_{B\to -I}(P) + \tau_{-I\to +I}(P). \tag{6}$$

Equation (6) makes the attribution problem explicit: the complete-method contrast $\tau_{B\to +I}(P)$ alone does not determine the matched intervention contrast $\tau_{-I\to +I}(P)$. The same complete-method change can be compatible with different contributions from the surrounding package and the claimed intervention.

We therefore use $\tau_{-I\to +I}(P)$ as the intervention-level attribution target. M^{-I} is obtained by the smallest method-side change that disables I—a zero loss weight or budget, or a bypassed stage—using the method's own switch where one exists, so that the non-intervention components of M^{+I} are preserved whenever they are scientifically separable from I. The resulting contrast is conditional on P and on this boundary: with interacting components it is a conditional marginal contribution rather than an order-invariant allocation of the complete-method change, and it does not by itself identify the mechanism through which I acts. Method-specific boundaries and implementation provenance appear in Appendix A.

3.3 Protocol-Conditional Attribution

We treat intervention attribution as conditional on a protocol P, which collects the non-intervention choices that govern how a model is trained, selected, and evaluated: preprocessing, optimization, training horizon, checkpoint selection, decision rule, evaluation implementation, and stochastic execution. The intervention state is not part of P; it is represented by the matched pair M^{-I} and M^{+I}. This separates two forms of variation. For a fixed P, variation over $\omega = (s, r)$ induces stochastic uncertainty in the estimated contrast; across scientifically defensible protocols, the expected contrast itself may differ, which we call *protocol dependence*. The resulting uncertainty about an intervention-level conclusion is distinct from stochastic uncertainty and is not treated as a probability distribution over contrasts. For two protocols P_a and P_b, or analogously two method configurations under the same protocol, define the attribution shift

$$\Delta_{\text{attr}}(P_a, P_b) = \tau_{-I\to +I}(P_b) - \tau_{-I\to +I}(P_a). \tag{7}$$

A nonzero Δ_{attr} indicates protocol dependence but does not, by itself, identify which protocol component generated the shift.

3.4 Decomposing Attribution Shifts

The aggregate shift in Equation (7) does not identify which protocol factor is associated with the change when several components of P vary simultaneously. Whenever possible, we therefore construct matched intermediate protocols. If two protocols differ only in a factor q, define

$$\Delta_q = \tau_{-I\to +I}(P^{(1)}) - \tau_{-I\to +I}(P^{(0)}). \tag{8}$$

When two binary factors vary jointly, we use a 2×2 factorial decomposition; the construction used for NIFTY/German is given in Appendix J.2.

A particularly important case arises when extending the training horizon also expands the checkpoint support available to model selection. Let $H_s < H_\ell$ denote short and long horizons. Let σ_s select from an independently trained short trajectory, σ_c select from the stored long trajectory with checkpoints restricted to those available by H_s, and σ_f select from the full long trajectory. Define

$$\psi(H, \sigma) = \tau_{-I \to +I}(P_{H,\sigma}). \tag{9}$$

Then the short-to-long attribution shift decomposes exactly as

$$\begin{aligned} \Delta_{\text{long}} &= \psi(H_\ell, \sigma_f) - \psi(H_s, \sigma_s) \\ &= \underbrace{\psi(H_\ell, \sigma_f) - \psi(H_\ell, \sigma_c)}_{S_{\text{sel}}} + \underbrace{\psi(H_\ell, \sigma_c) - \psi(H_s, \sigma_s)}_{R_{\text{traj}}}. \end{aligned} \tag{10}$$

The first term, S_{sel}, is a same-trajectory contrast associated with expanding checkpoint-selection support. The second, R_{traj}, retains differences between the independently executed short trajectory and the capped prefix of the long trajectory and is therefore treated as a trajectory residual rather than as an isolated training-horizon effect. For intervention state $z \in \{-I, +I\}$, define

$$E_z = \mathbb{E}_\omega \left[Y_\omega(M^z; P_{H_\ell, \sigma_f}) - Y_\omega(M^z; P_{H_\ell, \sigma_c}) \right]. \tag{11}$$

By construction, $S_{\text{sel}} = E_{+I} - E_{-I}$. These arm-level terms localize the selection-support shift but are diagnostic contrasts rather than separate intervention effects.

4 Experiments

We examine **RQ2** through three questions: whether complete-method evidence determines the matched intervention contribution (**RQ2a**, Section 4.2.1), how that contribution varies across method configurations and evaluation protocols (**RQ2b**, Section 4.2.2), and which training, selection, or component-level differences are associated with selected shifts (**RQ2c**, Section 4.2.3).

4.1 Experimental Design

Coverage. We evaluate eight settings: German (Asuncion & Newman, 2007), Bail (Jordan & Freiburger, 2015), Credit (Yeh & Lien, 2009), Income (Agarwal et al., 2021), and four Pokec-z/Pokec-n settings (Takac & Zabovsky, 2012) using region or gender as the sensitive attribute. The primary evaluation contains 36 supported cells across eleven methods, grouped descriptively as Modification (BIND, EDITS, FairEdit), Constraint (FairGNN, FairVGNN, SFG), and Regularization (BeMap, FairGB, FairSIN, GEAR, NIFTY). Coverage is intentionally non-Cartesian, so reported fractions describe the audited cells rather than prevalence across fairness-aware GNNs; coverage and exclusion criteria are given in Appendix B.

Controlled Protocol and Uncertainty. Each cell uses six train–validation–test splits with five stochastic runs nested within each split; B, M^{-I}, and M^{+I} share the split and run index of each of the 30 realizations, and M^{-I} and M^{+I} use identical seeds. Epoch-based methods are trained for $H = 200$ epochs with validation-BCE checkpoint selection, retaining method-specific components outside the designated intervention. We report ΔAUC, $-\Delta_{\text{DP}}$, and $-\Delta_{\text{EO}}$ (Appendix D.1) with 10,000 paired hierarchical-bootstrap replicates. A contrast is *resolved* when at least 75% of the matched units share the sign of the estimate, $|\hat{\tau}| \geq 0.01$, and the 95% percentile interval excludes zero; this is an operational reporting category rather than a hypothesis test (Appendices C and D).

External baseline. B is a vanilla GCN configured from a released command where one exists (German, including its 1000-epoch horizon) and from harness defaults with $H = 200$ otherwise; it is not matched to the capacity or backbone of each method, so $\tau_{B \to -I}$ absorbs these differences (Appendix C).

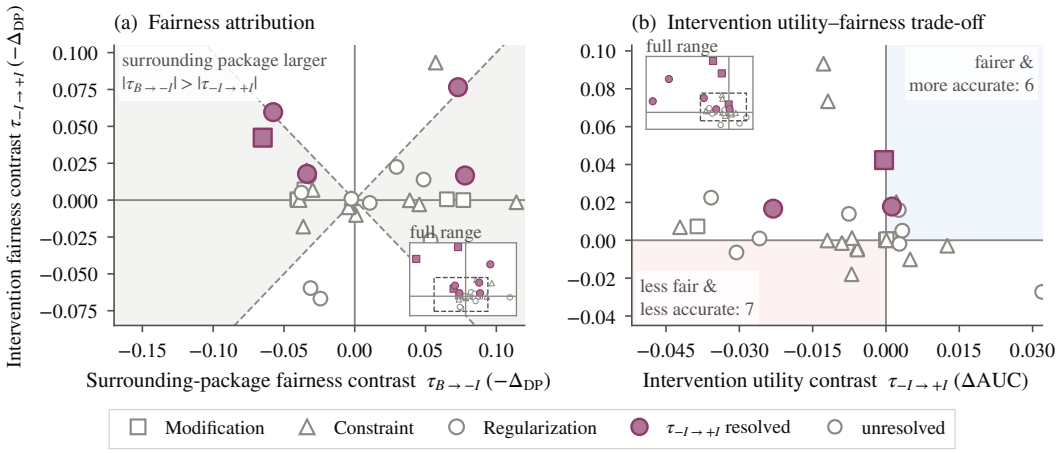


Figure 1: **Fairness attribution and intervention-level utility–fairness trade-offs.** (a) Surrounding-package versus matched intervention contrasts on $-\Delta_{\mathrm{DP}}$; the shaded region marks cells with a larger surrounding-package contrast. (b) Matched intervention contrasts in the $(\Delta\mathrm{AUC}, -\Delta_{\mathrm{DP}})$ plane; quadrant counts use point estimates without the $|\hat{\tau}| \geq 0.01$ threshold. Filled markers are resolved; insets show the full range.

Variations and targeted analyses. *Configuration variation* changes a released method-side choice under the controlled protocol, yielding 26 paired comparisons. *Protocol variation* holds the configuration fixed and changes either checkpoint selection, the training horizon, or broader published procedures; validation-BCE selection is retained in the latter two families, and B remains the fixed controlled reference throughout (Appendices E and F). Targeted follow-ups, excluded from all aggregate counts, examine two protocol-sensitive cells (NIFTY/German and FairGB/Bail; a planned FairGB/Credit companion was stopped by a pre-specified gate) and the component path of FMP (Appendices J and K). Task-adapted FnRGNN cells are reported separately (Appendix G.3).

4.2 Results

4.2.1 Complete-Method and Intervention-Level Attribution (RQ2a)

Package versus intervention. Figure 1(a) compares the surrounding-package contrast $\tau_{B\to -I}(P)$ with the matched intervention contrast $\tau_{-I\to +I}(P)$ on $-\Delta_{\mathrm{DP}}$. The surrounding-package contrast is larger in absolute magnitude in 27/36 cells for $\Delta\mathrm{AUC}$, 26/36 for $-\Delta_{\mathrm{DP}}$, and 29/36 for $-\Delta_{\mathrm{EO}}$, and the complete-method and intervention point estimates have opposite signs in 16/36, 18/36, and 15/36 cells, so a complete-method change does not by itself reveal the size or even the sign of the intervention's contribution (Table 1; cell-level estimates in Appendix G.1). Because $\tau_{B\to -I}$ depends on the external baseline, this comparison is sensitive to how strongly B is trained: with B trained for 1000 epochs on every dataset, the surrounding-package contrast is larger in 19/36 and 20/36 cells for $-\Delta_{\mathrm{DP}}$ and $-\Delta_{\mathrm{EO}}$, close to one half, and nearly all changed cells move from larger to not larger. Under the reported baseline, the pattern holds when restricting to directly separable intervention boundaries (10/12 and 11/12 cells) or excluding the two FairEdit cells whose intervention effect falls below a pre-specified re-execution threshold (24/34 and 27/34; Appendix G.2).

Size of intervention contributions. The matched intervention contrast is resolved in 11/36, 8/36, and 9/36 cells on the three coordinates. Where resolved, fairness contrasts almost always point in the intended direction: all 8 on $-\Delta_{\mathrm{DP}}$ and 8 of 9 on $-\Delta_{\mathrm{EO}}$ reduce disparity, with sizes from 0.014 to 0.276 (EDITS/German), and FairGB/Bail on $-\Delta_{\mathrm{EO}}$ is the only resolved increase. Unresolved does not mean absent: on $-\Delta_{\mathrm{DP}}$ and $-\Delta_{\mathrm{EO}}$, 21 of 28 and 21 of 27 unresolved contrasts have intervals extending beyond ± 0.01, whereas only 7 and 6 have intervals within ± 0.01, mostly for interventions that barely change the model (FairEdit and FairGNN/Bail). The evaluated interventions are therefore not ineffective; their matched contributions range from negligible to substantial. These counts, and all sensitivity analyses below, use only $\tau_{-I\to +I}$ and do not depend on B.

Table 1: **Family-level summary of package and intervention attribution across the 36 primary controlled cells.** Counts refer to method–dataset cells. "Resolved" indicates resolved intervention contrasts. "Package larger" indicates $|\tau_{B\to -I}| > |\tau_{-I\to +I}|$, and "Opposite sign" counts cells where $\tau_{B\to +I}$ and $\tau_{-I\to +I}$ have strictly opposite signs; cells with an exactly zero contrast are not counted. Family summaries are descriptive only.

	ΔAUC			$-\Delta_{\mathrm{DP}}$			$-\Delta_{\mathrm{EO}}$		
	Mod. (8)	**Const.** (13)	**Reg.** (15)	**Mod.** (8)	**Const.** (13)	**Reg.** (15)	**Mod.** (8)	**Const.** (13)	**Reg.** (15)
Resolved	1	2	8	3	0	5	3	0	6
Package larger	6	10	11	7	9	10	7	10	12
Opposite sign	4	7	5	5	6	7	4	6	5
Median $\lvert\tau_{B\to -I}\rvert$	0.107	0.024	0.046	0.053	0.036	0.038	0.041	0.017	0.030
Median $\lvert\tau_{-I\to +I}\rvert$	0.001	0.007	0.023	0.004	0.005	0.018	0.003	0.008	0.016

Utility–fairness direction. Figure 1(b) shows the utility–fairness direction associated with enabling the intervention itself. The most populated quadrant corresponds to lower DP disparity but lower AUC (16/36 cells), while the remaining cells span directions that improve both coordinates or worsen fairness despite changes in utility; requiring $|\hat{\tau}| \geq 0.01$ on both coordinates leaves 9 cells in the lower-disparity, lower-AUC quadrant and none in the quadrant that improves both. The intervention contribution is therefore heterogeneous in direction as well as magnitude.

4.2.2 Sensitivity of Intervention Attribution (RQ2b)

Method configuration. Across 26 paired alternative–primary configuration comparisons (Table 2), point-estimate signs change frequently on the fairness coordinates, but only 2 and 4 of these changes involve $|\hat{\tau}| \geq 0.01$ under both configurations, and no pair moves between resolved contrasts of opposite sign. Configuration changes instead shift magnitude: resolved status changes in 7–9 pairs per coordinate, and the paired shift Δ_{attr} has an interval excluding zero in 8–11 pairs, whereas none of the 15 differences between re-executions of five cells under identical settings does (Appendix D.5). For example, FairSIN/German changes from $+0.179$, resolved, on $-\Delta_{\mathrm{DP}}$ under GCN to $+0.001$ under GIN. Intervention attribution is therefore conditional on the evaluated configuration, with sensitivity expressed mainly through magnitude rather than direction (Appendices E and H).

Table 2: **Configuration sensitivity.** Counts over 26 paired alternative–primary comparisons; pair-level estimates appear in Appendix H.

Pairs with (of 26)	$-\Delta_{\mathrm{DP}}$	$-\Delta_{\mathrm{EO}}$	ΔAUC
Sign change	13	12	5
with $\lvert\hat{\tau}\rvert \geq 0.01$ in both	2	4	1
Resolved-status change	8	7	9
Δ_{attr} interval excludes 0	8	9	11

Evaluation protocol. Figure 2 and Table 3 summarize three protocol-sensitivity families that are analyzed separately. Checkpoint-selector replay changes the point-estimate sign in 10/36 primary cells and resolved status in 6/36, with either indicator changing in 14/36. The systematic published-horizon comparisons change sign in 1/16 cells and resolved status in none; among the three targeted published-horizon comparisons, one changes resolved status. In the four targeted published-procedure comparisons, three change sign, three change resolved status, and all four change at least one indicator. Among these sign changes, only one selector replay (FairGB/Credit) and one published-procedure comparison (FairGB/German) involve $|\hat{\tau}| \geq 0.01$ under both protocols. Complete paired results are reported in Appendix I. Because published-horizon and procedure comparisons use separate executions, we calibrate them against re-execution: cell-level contrasts on $-\Delta_{\mathrm{DP}}$ moved by up to 0.067 between

Table 3: **Summary of protocol-sensitivity outcomes.** Counts report changes in point-estimate sign, resolved status, or either within each comparison family.

Protocol comparison	Sign	Resolution	Either
Selector only	10/36	6/36	14/36
Published horizon, systematic	1/16	0/16	1/16
Published horizon, targeted	0/3	1/3	1/3
Published procedure	3/4	3/4	4/4

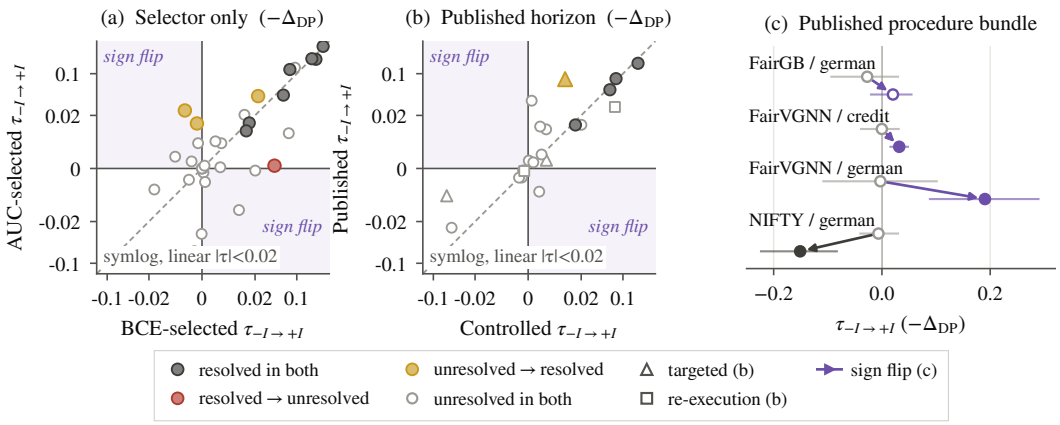


Figure 2: **Protocol sensitivity of the matched intervention contrast on $-\Delta_{\mathrm{DP}}$.** (a) Validation-BCE versus validation-AUC selection on the same trajectories; (b) controlled ($H = 200$) versus published horizons, with validation-BCE selection retained; (c) controlled versus published procedures for four targeted cells. Colors mark resolved-status transitions and shading marks sign changes; axes are symmetric-log, linear for $|\tau| < 0.02$.

identical re-executions, yet no re-execution difference had an interval excluding zero, whereas 3 of the 19 published-horizon comparisons do.

4.2.3 Diagnosing Attribution Shifts (RQ2c)

Targeted follow-ups, selected from the preceding analyses with procedures frozen before execution, examine whether shifts are associated with training, model selection, or the component path.

Longer training and checkpoint selection. NIFTY/German changes from -0.006, unresolved, under the controlled protocol to -0.151, resolved and negative, on $-\Delta_{\mathrm{DP}}$ under the published procedure, which restores the published horizon and drop rates while retaining validation-BCE selection. Because both the horizon H and the drop-rate setting D change, an $H \times D$ analysis first attributes most of this shift to the horizon: the horizon-associated contrast is resolved, at -0.117 averaged over D and -0.180 under the published D setting, whereas the drop-rate contrast is unresolved (Appendix J.2).

In the subsequent trajectory decomposition, the matched contrast changes from $+0.029$ at the short horizon to -0.148 under full long-trajectory selection; R_{traj} is near zero and $S_{\mathrm{sel}} = -0.177$. Figure 3 shows what this term contains: once checkpoints beyond epoch 200 become available, both arms are selected at later epochs, but at different ones (median 947 for M^{+I} and 692.5 for M^{-I}). Forcing both arms to a common checkpoint still yields -0.109, resolved, so most of the shift reflects the intervention's behavior at later training stages, and the remainder reflects the difference in selected stage (Appendix J.5).

FairGB/Bail provides a second-method case in which the published protocol change is restricted to training horizon. Its $-\Delta_{\mathrm{DP}}$ contrast changes from approximately -0.060 to -0.019, with $S_{\mathrm{sel}} = +0.047$ and $R_{\mathrm{traj}} = -0.006$; unlike NIFTY, the shift is positive and is associated primarily with the M^{+I} arm. Here the gap in selected checkpoints is larger (median 999 for M^{+I} and 63 for M^{-I}).

Component-level attribution. FMP exposes a staged path from a method-specific base state through propagation to an explicit fairness correction. Under the last-checkpoint selector, none of the 24 explicit fairness-correction contrasts across eight configurations and three coordinates is resolved, with all absolute estimates below 0.005; across the last, validation-BCE, and validation-AUC selectors, none of 72 evaluations is resolved. Larger changes in the complete method can instead arise at earlier stages of the path. For example, on Pokec-z with $\lambda_2 = 0.01$, the base transition is approximately $+0.026$, resolved, on $-\Delta_{\mathrm{DP}}$, whereas the subsequent explicit fairness-correction increments are approximately -0.002. Complete component and selector analyses are reported in Appendix K.

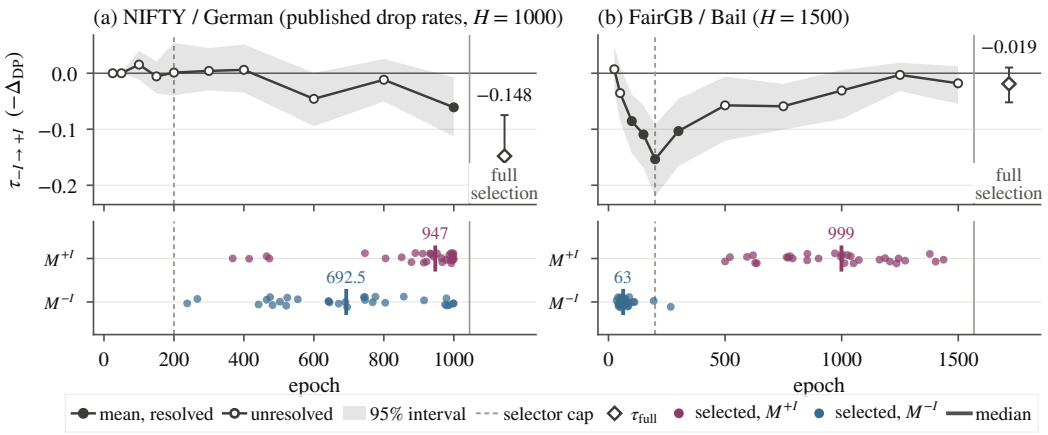


Figure 3: **Checkpoint-selection support in the targeted cases.** Top: matched contrast on $-\Delta_{\mathrm{DP}}$ at fixed checkpoints (mean over 30 units; shading, 95% intervals; dashed line, $H = 200$ selector cap) and under full-trajectory selection (τ_{full}, right). Bottom: epochs selected by validation BCE for M^{+I} and M^{-I}; bars mark medians.

5 Discussion

The results separate two kinds of evidence. Conclusions based on the intervention contrast alone, namely its heterogeneity, the resolved counts, and its sensitivity to configuration and checkpoint selection, do not depend on the external baseline, whereas the comparison with the surrounding package does: it holds in a majority of cells under the reported baseline but falls to about one half when the baseline is trained longer. Unresolved contrasts mostly reflect limited precision rather than absent effects, and the resolved counts change only modestly with the magnitude threshold (6 to 9 of 36 on $-\Delta_{\mathrm{DP}}$ for thresholds from 0.005 to 0.02; Appendix D.3). Sensitivity also differs across protocol families: checkpoint selection and procedure-level changes moved attribution, whereas restoring the published horizons of the systematic family rarely did.

Limitations. Our conclusions are limited by non-Cartesian coverage (Section 4.1), reliance on released implementations, binary node classification, the evaluated DP and EO criteria, a finite set of specifications, and 30 realizations per cell, which leave many fairness contrasts inconclusive at the 0.01 scale. Because the two arms select checkpoints independently, the matched contrast can include differences in selected training stage; the harness records test outputs only at the checkpoints chosen by the two selectors, so a common-checkpoint check was possible only for the targeted cases, where it reduced magnitudes without changing signs. The matched contrasts and decompositions are conditional on the evaluated specification and do not by themselves identify mechanisms, and the targeted analyses are diagnostic rather than representative.

6 Conclusion

We introduced protocol-conditional attribution, which separates complete-method performance from the incremental contribution of a claimed fairness intervention. Across controlled, configuration, and protocol analyses, complete-method and intervention-level evidence can diverge, and intervention attribution varies with the evaluated specification. Complete-method comparisons establish what a system does, whereas claims about why it is fairer require matched intervention-level evidence. Concretely, we recommend that a claim about a fairness intervention report (i) the matched contrast between the intervention-disabled and intervention-enabled versions alongside the complete-method comparison, (ii) the training horizon and checkpoint-selection rule under which it was obtained, and (iii) whether its sign and resolution persist under at least one alternative configuration or protocol. The same matched, protocol-conditional evaluation applies wherever an improvement is credited to a designated component, and extending it to other tasks, fairness criteria, and model families is a natural next step.

ETHICS STATEMENT

This work evaluates existing methods on public graph-fairness benchmarks; no new data were collected and no human subjects were involved. Sensitive attributes are used only as provided by the benchmarks to compute group disparities. Lower disparity under the evaluated criteria does not establish that a system is fair in deployment, and our results should not be used to certify any method as fair. Third-party code without an explicit license is not redistributed; the artifact retrieves it from the original repositories.

REPRODUCIBILITY STATEMENT

Implementation, protocol, and uncertainty details are in Appendices A–D. The anonymized artifact provides code, configurations, and per-unit metrics from which all reported numbers can be recomputed without retraining. Because most implementations use nondeterministic GPU kernels, retraining does not reproduce stored values exactly; re-executing five cells changed cell-level contrasts by up to 0.067, but no re-execution difference had an interval excluding zero (Appendix C).

AI USE STATEMENT

We used generative AI tools as coding assistants for evaluation, analysis, and figure scripts, and for language editing and manuscript organization; we did not use them to generate research ideas or interpret results. All AI-assisted code was verified against the frozen per-unit results, and all AI-assisted text was reviewed by the authors, who take full responsibility for the content of this work.

REFERENCES

Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *UAI*, pp. 2114–2124, 2021.

Arthur Asuncion and David Newman. UCI Machine Learning Repository, 2007.

Heloísa Oss Boll, Ali Amirahmadi, Mirfarid Musavian Ghazani, Wagner Ourique de Morais, Edison Pignaton de Freitas, Amira Soliman, Farzaneh Etminani, Stefan Byttner, and Mariana Recamonde-Mendoza. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *Journal of Biomedical Informatics*, 151:104616, 2024.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *MLSys*, 2021.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Qiang Chen, Zhongze Wu, Xiu Su, Xi Lin, Zhe Qu, Shan You, Shuo Yang, and Chang Xu. Stable fair graph representation learning with lipschitz constraint. In *ICML*, pp. 9354–9376, 2025.

André F. Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. In *ICLR*, 2024.

Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *WSDM*, pp. 680–688, 2021.

Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. Edits: Modeling and mitigating data bias for graph neural networks. In *WWW*, pp. 1259–1269, 2022a.

Yushun Dong, Song Wang, Yu Wang, Tyler Derr, and Jundong Li. On structural explanation of bias in graph neural networks. In *KDD*, pp. 316–326, 2022b.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023a.

Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li. Interpreting unfairness in graph neural networks via training node attribution. In *AAAI*, pp. 7441–7449, 2023b.

Yushun Dong, Zhenyu Lei, Zaiyi Zheng, Song Wang, Jing Ma, Alex Jing Huang, Chen Chen, and Jundong Li. Pygdebias: A python library for debiasing in graph learning. In *WWW*, pp. 1019–1022, 2024.

Yushun Dong, Song Wang, Zhenyu Lei, Zaiyi Zheng, Jing Ma, Chen Chen, and Jundong Li. Fairness-aware graph learning: A benchmark. In *KDD*, pp. 5402–5412, 2025.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods. In *ICLR*, 2024.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NIPS*, pp. 3315–3323, 2016.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Zirui Liu, Na Zou, Ali Mostafavi, and Xia Hu. Chasing fairness in graphs: A gnn architecture perspective. In *AAAI*, pp. 21214–21222, 2024.

Kareem L Jordan and Tina L Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3): 179–196, 2015.

Charlotte Laclau, Christine Largeron, and Manvi Choudhary. A survey on fairness for machine learning on graphs. *arXiv preprint arXiv:2205.05396*, 2022.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *NeurIPS*, pp. 109233–109282, 2024.

Zhixun Li, Yushun Dong, Qiang Liu, and Jeffrey Xu Yu. Rethinking fair graph neural networks from re-balancing. In *KDD*, pp. 1736–1745, 2024.

Xiao Lin, Jian Kang, Weilin Cong, and Hanghang Tong. Bemap: Balanced message passing for fair graph neural network. In *LoG*, pp. 37:1–37:18, 2024.

Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, and Na Zou. Learning fair graph representations via automated data augmentations. In *ICLR*, 2023.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: a gnn-based imbalanced learning approach for fraud detection. In *WWW*, pp. 3168–3177, 2021.

Donald Loveland, Jiayi Pan, Aaresh Farrokh Bhathena, and Yiyang Lu. Fairedit: Preserving fairness in graph neural networks through greedy graph editing. *arXiv preprint arXiv:2201.03681*, 2022.

Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. In *WSDM*, pp. 695–703, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

Soyoung Park and Sungsu Lim. Fnrgnn: Distribution-aware fairness in graph neural network. In *CIKM*, pp. 5110–5114, 2025.

Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. Fairsna: Algorithmic fairness in social network analysis. *ACM Computing Surveys*, 56(8):1–45, 2024.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. doi: 10.1038/s41562-020-0912-z.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International Scientific Conference & International Workshop Present Day Trends of Innovations*, 2012.

Xuwei Tan, Ziyu Hu, and Xueru Zhang. Benchmarking bias mitigation toward fairness without harm from vision to LVLMs. In *ICLR*, 2026.

Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *KDD*, pp. 1938–1948, 2022.

Cheng Yang, Jixi Liu, Yunhe Yan, and Chuan Shi. Fairsin: Achieving fairness in graph neural networks through sensitive information neutralization. In *AAAI*, pp. 9241–9249, 2024.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009.

A MATCHED INTERVENTION BOUNDARIES AND IMPLEMENTATION PROVENANCE

This appendix documents the method-specific constructions underlying the matched intervention contrast in Section 3.2. For each method, we identify the fairness intervention I represented by the evaluated implementation and construct M^{-I} by disabling that intervention while retaining the surrounding package whenever the components are scientifically separable. These constructions define the intervention boundary used in the empirical analysis; they do not introduce additional experimental results. Table 4 summarizes the notation used in the main text and appendices.

When a published fairness mechanism consists of coupled components, I denotes the corresponding composite intervention. The resulting matched contrast is attributed to that composite boundary and is not interpreted as identifying the contribution of its individual subcomponents unless a separate component-level analysis is explicitly constructed.

Table 4: Summary of notation.

Symbol	Meaning
M^{+I}	Evaluated method with its claimed fairness intervention I enabled
M^{-I}	The same method with I disabled; separable non-intervention components retained
B	external baseline (vanilla GCN)
P	Training, model-selection, and evaluation protocol, excluding the intervention state
$\omega = (s, r)$	Realization: data split s and run r nested within s
$\tau_{B\to -I},\ \tau_{-I\to +I},\ \tau_{B\to +I}$	Surrounding-package, matched intervention, and complete-method contrasts (Section 3.2)
$\Delta_{\mathrm{attr}}(P_a, P_b)$	Attribution shift between two specifications (Eq. (7))
$S_{\mathrm{sel}},\ R_{\mathrm{traj}}$	Selection-support shift and trajectory residual (Eq. (10))
$\Delta_{\mathrm{DP}},\ \Delta_{\mathrm{EO}}$	Demographic-parity and equal-opportunity disparities (Appendix D.1)
Complete method	M^{+I}, the released method with its claimed intervention enabled
Surrounding package	Components of M^{+I} other than I
Specification	A method configuration together with a training, selection, and evaluation protocol
Cell / resolved	A method–dataset pair / a contrast meeting the criterion in Section 4.1

A.1 PRIMARY MATCHED INTERVENTION CONSTRUCTIONS

Table 5 summarizes the intervention boundary for each of the eleven methods in the primary controlled evaluation. For a robustness analysis reported in Appendix G.2, we additionally distinguish directly separable intervention boundaries from composite boundaries. This classification is based only on the constructions in Table 5, before examining the resulting intervention contrasts. A boundary is treated as directly separable when enabling I changes a single designated fairness operation or objective rather than a coupled set of fairness-specific components.

Construction types. The eleven constructions in Table 5 fall into three descriptive types. (i) *Zero weight or budget*: the intervention's loss weight, budget, or strength is set to zero (BIND, $k = 0$; FairEdit, `edit_num`$= 0$; FairGNN, $\alpha = \beta = 0$; GEAR and NIFTY, `sim_coeff`$= 0$; FairSIN, $\delta = 0$ with the discriminator disabled). (ii) *Disabled stage or branch*: a designated stage or branch is switched off while the rest of the pipeline is unchanged (EDITS, bypassed debiasing stage; BeMap, full-graph training branch; FairGB, CAL and CNM disabled; FairVGNN, feature masking and weight clipping disabled). (iii) *Ablation-ladder row*: M^{-I} is the authors' own ablation row that precedes the intervention (SFG). These types describe how M^{-I} is obtained; they do not imply that M^{-I} reduces to an unconstrained backbone. Two constructions change more than a single operation: the SFG boundary also changes the generator objective weight (`loss_alpha`) following the authors' ablation ladder, and the BeMap boundary replaces per-epoch balanced subgraph sampling with full-graph training; results for the single-operation subset in Appendix G.2 should be read with this in mind.

A.2 CONFIGURATION PROVENANCE AND IMPLEMENTATION QUALIFICATIONS

The public repository associated with a method does not by itself determine the exact configuration evaluated in our study. Table 6 therefore records the repository file, released script, stored configuration, or benchmark artifact used to instantiate each primary setting, together with qualifications

Table 5: **Method-specific matched intervention constructions.** M^{-I} disables the claimed fairness intervention while retaining the scientifically separable surrounding package, whereas M^{+I} enables the intervention under the evaluated configuration.

Method	Backbone	Claimed intervention I	M^{-I}	M^{+I}
BIND	GCN	Influence-guided deletion of training nodes	No deletion ($k = 0$) under the same Stage-A model and retraining procedure	The highest-ranked harmful training nodes are deleted using the configured budget; the primary configuration uses $k = \mathrm{round}(0.01\|\mathcal{D}_{\mathrm{train}}\|)$
BeMap	GCN	Balance-aware per-epoch subgraph sampling	The method's own full-graph training branch, `train(-1, ...)`	A balance-aware fair subgraph is sampled at each training epoch, `train(epoch, ...)`
EDITS	GCN	Attribute and structural debiasing in the `fit()` stage	The debiasing stage is bypassed; the original features and adjacency are passed to the same prediction stage	The configured attribute- and structural-debiasing stage is applied before prediction
FairEdit	GCN	Gradient-guided edge editing; in the released configuration only the deletion path is active	`edit_num`=0	`edit_num`=10, deleting 10 undirected edges per run on each dataset
FairGB	SAGE	Contribution alignment (CAL) and counterfactual node mixup (CNM)	`use_cal=False`, `use_cnm=False`	CAL and CNM enabled under the dataset-specific configuration
FairGNN	GCN	Covariance fairness constraint and adversarial loss against a sensitive-attribute estimator	$\alpha = 0,\ \beta = 0$	Dataset-specific configured values of α and β
FairSIN	GCN	Feature neutralisation $x + \delta\,\mathrm{MLP}(x)$ and, where configured, a discriminator step	$\delta = 0$ and the discriminator disabled	Dataset-specific δ and discriminator configuration
FairVGNN	GCN	Generative feature masking and weight clipping	`f_mask=no`, `weight_clip=no`	Feature masking and weight clipping enabled under the evaluated configuration
GEAR	SAGE	Counterfactual-consistency similarity objective	`sim_coeff`=0	`sim_coeff`=0.6
NIFTY	GCN	Counterfactual- and noise-invariance objective on the constrained encoder	`sim_coeff`=0	The configured invariance objective is enabled
SFG	SAGE	Lipschitz constraint on the encoder together with the stronger χ^2-DRO generator objective	No constraint and `loss_alpha`=1, corresponding to the authors' unconstrained FairVGNN row	Constraint enabled with $\rho = 2.0$ and `loss_alpha`=0.5

relevant to interpreting $\tau_{-I\to+I}(P)$. These qualifications were fixed during the implementation audit rather than introduced in response to the observed intervention contrasts.

Repository versioning. Repository URLs identify the public upstream implementations, while Table 6 identifies the concrete configuration sources used in the evaluation. An exact commit is reported only when the evaluated revision can be recovered from the frozen experimental artifact; we do not substitute a current upstream revision for an unrecorded evaluated revision.

A.3 Task-Adapted FnRGNN Construction

FnRGNN[1] is reported separately from the primary controlled evaluation because its released implementation targets node regression rather than the node-classification setting used in the main study. Its released model class and stored best configurations are retained, while the evaluation harness completes the task-specific classification loop.

[1] `https://github.com/sybeam27/FnRGNN`

Table 6: **Configuration provenance and implementation qualifications.** The source column identifies the artifact used to instantiate each primary matched evaluation.

Method	Evaluated configuration source	Qualification
BIND	`BIND-main` parser defaults and the documented released procedure	The released influence estimator originally evaluates its fairness cost using test labels. For the controlled evaluation, this cost is evaluated on validation nodes to remove test leakage while leaving the remainder of the estimator unchanged. The 1% deletion budget is the primary configuration; the larger budget is analyzed separately as a configuration variation.
BeMap	`BeMap-main/README.md` command and `train_bemap.py` parser defaults	The primary construction uses the method's own full-graph and fair-subgraph branches. The released model-selection procedure accesses the test split and is therefore not used as an admissible native evaluation.
EDITS	Stored `utils/param.json` values and the evaluation wrapper	The wrapper's parameter routing did not activate the stored debiasing settings as intended. We repair the routing according to the stored configuration values; no experimental output is modified after evaluation.
FairEdit	Stored `utils/param.json` values and the evaluation wrapper	The repository-default intervention deletes 10 undirected edges per run on every evaluated dataset, i.e., 0.045%, 0.0031%, and 0.0007% of the edges of German, Bail, and Credit, respectively; the released edge-addition branch is disabled in the repository configuration. Small $\tau_{-I \to +I}(P)$ values should therefore be interpreted together with this intervention magnitude.
FairGB	Released `run.sh` configurations	Dataset-specific CAL/CNM parameters are taken from the released scripts. Parameters not explicitly set by the scripts retain the corresponding repository defaults.
FairGNN	Official Pokec scripts where available; benchmark configuration files for the remaining datasets	The original repository directly provides configurations for its released Pokec/NBA settings but not for German, Bail, or Credit. The latter therefore have weaker first-party configuration provenance.
FairSIN	Released ablation commands in `experiment.sh` with `in-train.py`	The designated GCN configuration is used for the primary evaluation. Additional GIN and SAGE configurations are analyzed as configuration variations rather than additional primary cells.
FairVGNN	Released dataset-specific run scripts and corresponding stored provenance artifacts	The published ablation rows may change optimization parameters together with intervention flags. In the primary matched comparison, the surrounding configuration is held fixed between M^{-I} and M^{+I}; published alternative rows are analyzed separately as configuration variations.
GEAR	Released `src/main.py` defaults and released counterfactual assets	The released assets correspond to sensitive-attribute flips on the unchanged graph and do not reproduce the complete counterfactual graph-generation procedure described by the method. The reported contrast is therefore interpreted as the contribution of the similarity objective under the released assets.
NIFTY	Released German configuration and stored benchmark configurations for Bail and Credit	The public repository directly documents the German experiment. Bail and Credit are supported by the evaluation implementation but have weaker published-command provenance.
SFG	Released `run.sh` ablation ladder with `sfg.py` and `sfg_credit.py`	The matched construction follows the authors' ablation ladder: M^{-I} corresponds to the unconstrained FairVGNN row, whereas M^{+I} adds the encoder constraint and its associated stronger generator objective.

Table 7: **Task-adapted FnRGNN matched construction.** The three cells are reported separately from the primary controlled evaluation.

Backbone	Claimed intervention I	M^{-I}	M^{+I}
GCN	Edge reweighting, MMD representation alignment, and GWN prediction normalisation	The three fairness components are disabled while retaining the same task-adapted GCN training loop	Edge reweighting, MMD alignment, and GWN normalisation are enabled

Specifically, the regression MSE objective is replaced with binary cross-entropy, losses are restricted to training nodes, and the missing `mmd_sample` and training-horizon settings are supplied by the harness. These task adaptations are shared by M^{-I} and M^{+I} and therefore do not define the intervention contrast.

The evaluated task-adapted configurations are taken from the released best-configuration files for German, Pokec-z, and Pokec-n. Because completing the node-classification training loop requires task-level adaptation beyond the released node-regression procedure, these cells are never pooled with the primary controlled evaluation. FMP is not included in Table 5 because its implementation supports a more informative component path than a single binary intervention switch; its base, propagation, and explicit fairness-correction stages are analyzed separately in Appendix K.

B Evaluation Coverage and Exclusions

Table 8: **Evaluation sets used in the binary-intervention analyses.** Counts refer to completed cells, each containing 30 matched split–run realizations. Rows answer distinct scientific questions and are not pooled into a common frequency.

Evaluation set	Cells	Role
Primary controlled	36	Main attribution analysis
Task-adapted controlled	3	FnRGNN, reported separately
Configuration robustness	26	Method-configuration sensitivity
Systematic published horizon	16	Systematic protocol comparison
Targeted published horizon	3	Targeted protocol comparison
Targeted published procedure	4	Broader procedure comparison
Re-execution pairs	2	SFG pairs with unchanged horizon

The evaluation is partitioned into sets that answer different scientific questions and are therefore not pooled into a common aggregate. Table 8 summarizes the completed binary-intervention evaluation cells. The 36 primary controlled cells support the main package-versus-intervention analysis; task-adapted FnRGNN cells, configuration variations, and protocol variations are reported separately. Selector replays reuse stored controlled trajectories and therefore do not constitute additional trained evaluation cells.

The FMP study is not included in Table 8 because it uses a multi-stage component path rather than a single binary intervention boundary; it is analyzed separately in Appendix K.

B.1 Primary Controlled Coverage

Table 9 reports the 36 method–dataset settings included in the primary controlled evaluation. A check mark denotes one designated primary configuration for which the released implementation, required artifacts, and a scientifically defensible matched M^{-I}/M^{+I} construction were available. Blank entries are not failed experiments or negative results.

Coverage is intentionally non-Cartesian. A method–dataset setting is included only when the released implementation, required artifacts, and intervention construction support a valid matched comparison. We do not tune new configurations solely to populate missing entries, and coverage decisions are made independently of the observed $\tau_{-I\to+I}(P)$. Consequently, aggregate fractions over the 36 primary cells are descriptive summaries of the audited cells and are not interpreted as prevalence estimates over all fairness-aware GNN methods or datasets.

Coverage audit. The accompanying anonymized artifact[2] provides a machine-readable coverage ledger, `harness/coverage_manifest.csv`, spanning 126 candidate method–dataset settings considered in the broader feasibility audit. This ledger is broader than the 36-cell primary evaluation and includes candidates considered for non-primary evaluation roles; it is therefore not used as the denominator for primary frequency summaries. For every excluded candidate, the manifest records one of four fixed exclusion categories—unsupported released configuration, absence of a scientifically separable intervention-disabled state, an invalid evaluation split, or a required artifact being unavailable—together with the corresponding frozen ledger note. The manifest is generated from the frozen audit records rather than reconstructed from the observed intervention results.

B.2 Separate Evaluation Sets and Completeness

The three task-adapted FnRGNN cells cover German, Pokec-z, and Pokec-n. They follow the same matched-intervention principle but are excluded from the primary aggregate because the released method targets node regression and requires task-level adaptation for the common node-classification evaluation; the construction is detailed in Appendix A.3.

Configuration-robustness cells retain the controlled evaluation protocol while changing a method-side configuration. Protocol-variation cells instead retain the evaluated method configuration while changing training, checkpoint selection, or broader published procedures. The 16 systematic published-horizon cells, three targeted published-horizon cells, four targeted published-procedure

[2] `https://anonymous.4open.science/r/FairGNN-Eval-ICLR`

Table 9: **Primary controlled evaluation coverage.** Pokec-z_R and Pokec-n_R use region as the sensitive attribute, whereas Pokec-z_G and Pokec-n_G use gender. The table contains 36 primary cells.

Method	German	Bail	Credit	Income	Pokec-z_R	Pokec-z_G	Pokec-n_R	Pokec-n_G
FairGNN	✓	✓	✓	–	✓	✓	✓	✓
NIFTY	✓	✓	✓	–	–	–	–	–
FairVGNN	✓	✓	✓	–	–	–	–	–
FairGB	✓	✓	✓	–	–	–	–	–
FairSIN	✓	✓	✓	–	✓	–	✓	–
EDITS	✓	✓	✓	–	–	–	–	–
FairEdit	✓	✓	✓	–	–	–	–	–
BeMap	–	✓	✓	–	✓	–	–	–
GEAR	–	✓	–	–	–	–	–	–
BIND	–	✓	–	✓	–	–	–	–
SFG	✓	✓	✓	–	–	–	–	–
Total	8	11	9	1	3	1	2	1

Table 10: **Common controlled-evaluation contract.** Runs are nested within data splits, and the 30 matched realizations in each cell are not treated as independent observations.

Component	Controlled setting
Data splits	6 train–validation–test partitions
Runs per split	5 stochastic runs nested within split
Matched unit	Same split and run index; identical seeds for M^{-I} and M^{+I}
Systems per unit	B, M^{-I}, and M^{+I}
Primary horizon	$H = 200$
Primary selector	Validation BCE
Selector robustness	Validation AUC replay
external baseline B	Vanilla GCN shared across methods (Appendix C)
Evaluator	Common within each matched comparison

cells, and two re-execution pairs are reported as separate families and are never combined into a single success rate.

All 90 binary-intervention cells represented in Table 8 are complete and contain the required 30 matched split–run realizations. Pending or partially completed cells are excluded from scientific summaries. The accompanying artifact additionally provides complete-method intervention and score-convention manifests, frozen paper-facing result tables, and a compact per-unit metric export from which the reported matched contrasts can be reconstructed.

C Controlled Evaluation Protocol

The primary evaluation uses a common controlled protocol while preserving method-specific components that are not part of the designated fairness intervention. Table 10 summarizes the resulting evaluation contract.

Nested matched replication. Each evaluation cell contains six train–validation–test splits and five stochastic runs within each split, yielding $6 \times 5 = 30$ matched realizations. For realization $\omega = (s, r)$, B, M^{-I}, and M^{+I} use the same split and run index, and M^{-I} and M^{+I} use identical seeds. The resulting 30 realizations are therefore a nested replication design rather than 30 independent observations: variation across splits reflects sensitivity to the sampled data partition, whereas repeated runs within a split reflect stochastic-training variation conditional on that partition. The same hierarchy and within-unit pairing are retained by the uncertainty procedure in Appendix D. Matching fixes the split, the run index, and the seeds of the two intervention arms but does not make training bitwise deterministic, because the sparse propagation kernels are nondeterministic on GPU. The resulting run-to-run variation enters the observed run-level variability and is therefore reflected in the hierarchical-bootstrap intervals; its magnitude is reported in the Reproducibility Statement.

Training contract and method-specific components. For epoch-based methods, the primary controlled horizon is $H = 200$. Method-specific architecture, optimization, preprocessing, and internal stages are retained unless they constitute the intervention being disabled; the relevant control is within the matched M^{-I}/M^{+I} pair rather than across different methods. When a released imple-

Table 11: **External baseline B per dataset.** All reported contrasts involving B use rep 1, one B per unit shared by every method, trained for the published horizon where the configuration resolver returns one (German) and $H = 200$ otherwise; rep 2 repeats this rule independently, and H1000 trains for 1000 epochs on every dataset. "Frozen" shows an earlier baseline, trained separately for each method (distinct draws given) and not used in any reported result, for transparency. AUC and Δ_{DP} are B's own test values at the validation-BCE checkpoint, averaged over 30 units.

	Frozen	Horizon			B test AUC			
Dataset	draws	rep 1	rep 2	H1000	frozen	rep 1	rep 2	H1000
german	9	1000	1000	1000	0.4415	0.6543	0.6544	0.6525
bail	11	200	200	1000	0.7734	0.7734	0.7734	0.9069
credit	9	200	200	1000	0.7186	0.7177	0.7179	0.7304
income	1	200	200	1000	0.6530	0.6520	0.6538	0.7027
pokec_z	4	200	200	1000	0.7038	0.7038	0.7038	0.7290
pokec_n	3	200	200	1000	0.7112	0.7112	0.7112	0.7145
pokec_z_g	1	200	200	1000	0.7038	0.7038	0.7038	0.7290
pokec_n_g	1	200	200	1000	0.7110	0.7110	0.7110	0.7144

Notes. German is the only dataset for which the resolver returns a published horizon (1000); elsewhere no published value exists and the controlled default of 200 stands, so *frozen* and *rep 1* differ only in that B is now one draw per unit rather than one per method. German's frozen baseline is below chance because that horizon was never read.

mentation uses a different training unit, the corresponding controlled budget is applied in that unit: GEAR uses 200 mini-batch steps, and BIND retains its Stage-A pretraining while the $H = 200$ budget applies to the subsequent retraining stage. Method-specific intervention boundaries are documented in Appendix A.

Checkpoint selection. The primary checkpoint is selected by validation binary cross-entropy over the available training trajectory; no test-set quantity is used for controlled checkpoint selection. Within each matched comparison, M^{-I} and M^{+I} use the same selection rule and evaluator. Validation-AUC selection is evaluated separately by replaying selection on the same stored trajectory, without retraining, and is therefore treated as protocol sensitivity rather than as part of the primary estimate.

Seeds and compute environment. Run $r \in \{0, \ldots, 4\}$ uses seed $27+r$ in every split. Within each method, M^{-I} and M^{+I} are initialized from the same seed; the rule by which each implementation derives its random state from this seed differs across implementations and is recorded for every unit in the artifact. Experiments were run on a single NVIDIA GPU with 48 GB of memory, with several processes sharing the device. Three Python 3.11 environments were used: two with PyTorch 2.6.0 (CUDA 12.4), differing in auxiliary packages, for most implementations, and one with PyTorch 2.2.2 (CUDA 12.1), PyTorch Geometric 2.5.3, and DGL 1.1.3 for the DGL-dependent implementations; package lists captured after the runs are included in the artifact. The execution intervals recorded in the run logs sum to approximately 52 GPU-hours. This is a lower bound: 12 primary cells (FairGNN, NIFTY, FairGB, and FairVGNN on German, Bail, and Credit) and several auxiliary runs have no timing records, and because processes shared the GPU, compute cannot be attributed to individual cells.

external baseline. The external baseline B is a vanilla GCN trained with the same harness, splits, run indices, and validation-BCE selection as the evaluated methods, and is trained once per split–run unit and shared by all methods. Its configuration follows the released NIFTY GCN command on German, including the 1000-epoch horizon, and harness defaults with $H = 200$ on the remaining datasets, for which no released command exists; this harness GCN has not been verified against an upstream reference implementation. Table 11 reports its configuration and test metrics, together with an independent retraining under the same rule and a variant trained for 1000 epochs on every dataset. Independent retraining changes no package-larger classification, whereas the 1000-epoch variant changes 7, 9, and 11 cells on $\Delta\mathrm{AUC}$, $-\Delta_{\mathrm{DP}}$, and $-\Delta_{\mathrm{EO}}$, almost all from larger to not larger (Table 12).

Table 12: **Robustness of the primary package-versus-intervention magnitude summary.** "Package larger" denotes $|\tau_{B\to -I}| > |\tau_{-I\to +I}|$. The upper block uses the reported external baseline; the lower block retrains B for 1000 epochs on every dataset. Method-balanced averages within-method proportions equally across the eleven methods; LOMO gives the range of cell-weighted proportions after excluding each method in turn. Subsets and exclusions are defined in the text.

Analysis	ΔAUC	$-\Delta_{\mathrm{DP}}$	$-\Delta_{\mathrm{EO}}$
Reported external baseline			
All primary cells	27/36	26/36	29/36
Single-operation subset	10/12	10/12	11/12
excluding FairEdit	7/9	7/9	8/9
Excluding inactive interventions	25/34	24/34	27/34
Method-balanced	78.6%	79.0%	85.2%
Leave-one-method-out range	72.7–78.8%	69.7–79.3%	78.8–86.2%
external baseline trained for 1000 epochs			
All primary cells	20/36	19/36	20/36
Method-balanced	62.1%	52.7%	55.7%
Leave-one-method-out range	51.5–60.6%	48.5–58.1%	51.5–61.3%

Table 13: **Re-execution differences in the cell-level mean intervention contrast.** Upper block: five primary cells re-executed twice under the frozen configuration on a single machine and code state, giving three realizations and three pairwise differences per cell. Lower block: the two SFG cells whose controlled and published-horizon runs differ only by re-execution, recorded separately in the original study. Entries give the difference and its 95% paired hierarchical-bootstrap interval; no interval excludes zero in either block. Re-execution variation differs by more than an order of magnitude across cells (on $-\Delta_{\mathrm{DP}}$, from 0.001 for NIFTY/German to 0.067 for SFG/German).

Cell	Pairing	ΔAUC	$-\Delta_{\mathrm{DP}}$	$-\Delta_{\mathrm{EO}}$
Re-executed for this audit				
NIFTY / German	frozen → rep 1	−0.0005 [−0.0019, 0.0004]	−0.0008 [−0.0034, 0.0022]	−0.0021 [−0.0074, 0.0023]
	frozen → rep 2	−0.0002 [−0.0009, 0.0004]	0.0004 [−0.0057, 0.0067]	0.0023 [−0.0039, 0.0081]
	rep 1 → rep 2	0.0003 [−0.0008, 0.0016]	0.0013 [−0.0048, 0.0082]	0.0044 [−0.0027, 0.0116]
FairVGNN / German	frozen → rep 1	0.0001 [0.0000, 0.0003]	0.0101 [−0.0014, 0.0418]	0.0083 [−0.0024, 0.0357]
	frozen → rep 2	0.0001 [0.0000, 0.0003]	0.0100 [−0.0020, 0.0418]	0.0090 [0.0000, 0.0358]
	rep 1 → rep 2	0.0000 [0.0000, 0.0000]	−0.0001 [−0.0006, 0.0000]	0.0006 [0.0000, 0.0025]
FairGB / Bail	frozen → rep 1	0.0005 [−0.0036, 0.0052]	−0.0062 [−0.0233, 0.0098]	−0.0078 [−0.0274, 0.0075]
	frozen → rep 2	−0.0020 [−0.0051, 0.0016]	−0.0008 [−0.0139, 0.0109]	0.0003 [−0.0151, 0.0131]
	rep 1 → rep 2	−0.0025 [−0.0065, 0.0010]	0.0054 [−0.0060, 0.0177]	0.0080 [−0.0043, 0.0217]
FairSIN / Credit	frozen → rep 1	0.0031 [−0.0202, 0.0300]	0.0113 [−0.0090, 0.0322]	0.0083 [−0.0098, 0.0273]
	frozen → rep 2	0.0117 [−0.0140, 0.0407]	−0.0073 [−0.0359, 0.0165]	−0.0075 [−0.0312, 0.0129]
	rep 1 → rep 2	0.0087 [−0.0088, 0.0260]	−0.0187 [−0.0465, 0.0036]	−0.0158 [−0.0388, 0.0033]
SFG / German	frozen → rep 1	0.0098 [−0.0228, 0.0428]	−0.0668 [−0.1461, 0.0027]	−0.0460 [−0.1200, 0.0174]
	frozen → rep 2	0.0086 [−0.0165, 0.0364]	−0.0179 [−0.1207, 0.0846]	−0.0052 [−0.0972, 0.0885]
	rep 1 → rep 2	−0.0012 [−0.0194, 0.0163]	0.0489 [−0.0696, 0.1713]	0.0408 [−0.0703, 0.1514]
Frozen pairs reclassified as re-execution				
SFG / German	controlled → published	−0.0032 [−0.0342, 0.0202]	−0.0457 [−0.1218, 0.0324]	−0.0298 [−0.0907, 0.0350]
SFG / Credit	controlled → published	−0.0003 [−0.0024, 0.0019]	0.0005 [−0.0082, 0.0128]	−0.0014 [−0.0087, 0.0070]

Notes. All entries use the same estimator as the main analysis: 30 matched units, splits resampled and then runs within each drawn split, 10,000 replicates, fixed seed. The upper block was produced for this audit; the lower block is two runs from the original study, so the blocks are not pooled.

Re-execution noise. To calibrate comparisons between separately executed runs, we re-executed five primary cells twice under the frozen configuration; the cells (FairGB/Bail, FairSIN/Credit, FairVGNN/German, NIFTY/German, and SFG/German) were chosen before re-execution because they also appear in the published-horizon or published-procedure comparisons. Table 13 reports the change in the cell-level mean intervention contrast and its paired hierarchical-bootstrap interval. None of the 45 differences (15 pairs $\times$ 3 coordinates) has an interval excluding zero, so the reported intervals cover re-execution variation. The magnitude of re-execution variation, however, differs by an order of magnitude across cells: the largest change on $-\Delta_{\mathrm{DP}}$ is 0.067 for SFG/German but 0.001 for NIFTY/German. Point-estimate-based counts, such as sign changes, should therefore be read together with the intervals, and we report re-execution variation per cell rather than as a single noise floor. The two SFG pairs whose published horizon equals the controlled horizon provide an addi-

tional, independently recorded re-execution comparison, reported as a separate block; none of their six intervals excludes zero, and the recorded SFG/German difference on $-\Delta_{\mathrm{DP}}$ (0.046) lies within the range produced by the deliberate re-executions (0.018–0.067).

D METRICS, SCORE CONVENTIONS, AND UNCERTAINTY

D.1 DECISION MARGINS AND EVALUATION METRICS

For every evaluated classifier, q_i denotes a canonical binary decision margin whose interpretation is fixed by the classifier's training objective and output parameterization, independently of the observed evaluation results. For classifiers trained with binary cross-entropy with logits, the scalar classifier logit is used directly, $q_i = \ell_i$. For two-class classifiers trained with cross-entropy, we use the logit margin $q_i = \ell_{i,1} - \ell_{i,0}$, for which $q_i > 0$ is exactly equivalent to $\mathrm{softmax}(\ell_i)_1 > 0.5$. The positive class is fixed as class 1, and score orientation is never selected from evaluation labels or outcomes.

Binary predictions are therefore defined uniformly as $\hat{y}_i = \mathbb{I}[q_i > 0]$. Under these conventions, zero represents the same group-independent binary decision threshold across the evaluated implementations rather than a common numerical threshold applied to heterogeneous raw score types. Predictive utility is measured by ROC–AUC using the same margins $\{q_i\}$. Because the corresponding class-1 probability is a strictly increasing function of the binary logit margin, this representation preserves the AUC ranking relative to the equivalent probability representation.

The binary-intervention evaluations use the common evaluation interface described in Appendix C. FMP is evaluated through its separate component-analysis path, but uses the same canonical margin and fairness definitions. The accompanying artifact records the implementation-level loss, raw output type, canonical margin, positive class, decision rule, and supporting code path for all fourteen implementations in `harness/score_convention_manifest.csv`.

For the binary sensitive attribute $z_i \in \{0, 1\}$, demographic-parity disparity is

$$\Delta_{\mathrm{DP}} = \left| \widehat{\Pr}(\hat{y} = 1 \mid z = 0) - \widehat{\Pr}(\hat{y} = 1 \mid z = 1) \right|,$$

and equal-opportunity disparity is

$$\Delta_{\mathrm{EO}} = \left| \widehat{\Pr}(\hat{y} = 1 \mid y = 1, z = 0) - \widehat{\Pr}(\hat{y} = 1 \mid y = 1, z = 1) \right|.$$

For reporting contrasts, all coordinates are oriented so that larger is better:

$$Y(M; P, \omega) = \left[\mathrm{AUC}(M; P, \omega), -\Delta_{\mathrm{DP}}(M; P, \omega), -\Delta_{\mathrm{EO}}(M; P, \omega) \right].$$

Thus, a positive fairness contrast corresponds to a reduction in the corresponding disparity. The three coordinates are reported separately and are not scalarized into a single utility–fairness objective.

D.2 PAIRED HIERARCHICAL BOOTSTRAP

Each completed cell contains six data splits and five stochastic runs nested within each split. For every outcome coordinate and matched contrast, uncertainty is estimated using 10,000 paired hierarchical-bootstrap replicates. In each replicate, we first sample six split identifiers with replacement; for every sampled split, we then sample five run identifiers with replacement from that split. The complete matched observation (B, M^{-I}, M^{+I}) is retained for every sampled split–run unit, and the corresponding mean contrast is recomputed.

The reported 95% uncertainty interval is the percentile interval defined by the 2.5th and 97.5th percentiles of the bootstrap distribution. This construction preserves both the split-level hierarchy and the within-unit pairing required by $\tau_{B \to -I}(P)$, $\tau_{-I \to +I}(P)$, and $\tau_{B \to +I}(P)$; the 30 matched realizations are therefore not treated as 30 independent top-level observations.

For auditability, the accompanying artifact provides the matched split–run metric export in `results/per_unit_metrics.csv.gz`. It contains the B, M^{-I}, and M^{+I} outcomes needed to reconstruct the reported contrasts and their hierarchical-bootstrap intervals without redistributing training data, model weights, or trajectory artifacts.

Table 14: **Classifier score conventions.** The canonical decision margin is fixed from the classifier's training objective and output parameterization rather than from evaluation performance. All binary predictions use $\hat{y}_i = \mathbb{I}[q_i > 0]$.

Implementation	Classifier output and loss	Canonical margin q_i
B, BIND, BeMap, EDITS, FairEdit, FairGB, FairGNN, FairSIN, FairVGNN, GEAR, NIFTY, SFG, FnRGNN	Scalar logit; binary cross-entropy with logits	$q_i = \ell_i$
FMP	Two-class logits; cross-entropy	$q_i = \ell_{i,1} - \ell_{i,0}$

Table 15: **Resolved counts under alternative magnitude thresholds.** Unit-level sign consistency ≥ 0.75 and a 95% interval excluding zero are required throughout; the reported threshold is 0.01.

Magnitude threshold	$\Delta\mathrm{AUC}$	$-\Delta_{\mathrm{DP}}$	$-\Delta_{\mathrm{EO}}$
$\lvert\hat{\tau}\rvert \geq 0.005$	15	9	9
$\lvert\hat{\tau}\rvert \geq 0.010$ (reported)	11	8	9
$\lvert\hat{\tau}\rvert \geq 0.020$	7	6	7

D.3 Resolution Criterion and Interpretation

For each estimated contrast, we additionally record *unit-level sign consistency*, defined as the fraction of matched split–run units with a nonzero contrast whose sign agrees with the sign of the point estimate; units with an exactly zero contrast are excluded from the denominator. An estimate is labeled *resolved* only when all three conditions hold:

$$\text{sign consistency} \geq 0.75, \qquad |\widehat{\tau}| \geq 0.010, \qquad 0 \notin \left[\mathrm{CI}^{2.5\%}, \mathrm{CI}^{97.5\%}\right].$$

The magnitude threshold prevents numerically small contrasts from being labeled resolved solely because of a narrow interval. The sign-consistency condition requires that the direction of the estimate be shared by at least three quarters of the matched units, so that a resolved contrast is not driven by a minority of large unit-level contrasts. Because it operates on matched units rather than on bootstrap replicates, this condition is not implied by the interval condition: 2, 4, and 2 primary cells on $\Delta\mathrm{AUC}$, $-\Delta_{\mathrm{DP}}$, and $-\Delta_{\mathrm{EO}}$ satisfy the interval and magnitude conditions but not sign consistency. Replacing it with a bootstrap-replicate sign criterion would label 13/36, 12/36, and 11/36 cells resolved; all point estimates and intervals are reported irrespective of the label.

Table 15 varies the magnitude threshold while keeping the other two conditions fixed. The resolved counts change modestly, and in every case only a minority of intervention contrasts is resolved; on $-\Delta_{\mathrm{EO}}$, the count is identical at 0.005 and 0.01, so the interval condition rather than the magnitude threshold is binding.

Although exclusion of zero from the interval is one component of this rule, *resolved* is used as an operational reporting category rather than as a null-hypothesis significance test. We do not interpret the collection of resolved cells as a family of statistical discoveries. Accordingly, aggregate resolved counts and sign-change counts are descriptive summaries of the audited cells, not familywise- or false-discovery-rate-adjusted discovery rates.

Because the primary hierarchy contains only six top-level split clusters, we additionally assess the sensitivity of the resolved classification to an alternative split-level interval and to leave-one-split-out perturbations in Appendix D.4.

D.4 Six-Split Sensitivity Analysis

The primary uncertainty procedure resamples six top-level data splits and then the five matched runs nested within each sampled split. Because the number of top-level clusters is small, we additionally evaluate sensitivity to an alternative split-level interval construction. For each primary cell and outcome coordinate, we first average the five matched run-level contrasts within each split, yielding six split-level estimates, and compute

$$\bar{\tau} \pm t_{0.975,5} \frac{s_{\mathrm{split}}}{\sqrt{6}}.$$

Table 16: **Sensitivity of the resolved classification to split-level inference.** "Split-level t" averages the matched contrasts within each of the six splits and replaces the primary percentile interval with the $t_{0.975,5}$ interval computed from these six split-level estimates; the magnitude threshold is unchanged, and the sign condition is applied to the six split-level estimates rather than to the 30 matched units. "Agreement" is the number of primary cells whose resolved/unresolved label is unchanged.

Coordinate	Primary	Split-level t	Primary resolved lost	Agreement
$\Delta\mathrm{AUC}$	11/36	13/36	0	34/36
$-\Delta_{\mathrm{DP}}$	8/36	10/36	0	34/36
$-\Delta_{\mathrm{EO}}$	9/36	11/36	0	34/36

This sensitivity analysis changes the interval construction rather than the estimated contrast: the point estimates agree with the frozen primary estimates to numerical precision.

No primary resolved cell becomes unresolved under the six-split t-interval analysis on any coordinate. Two additional cells are classified as resolved on each coordinate under this alternative interval, giving resolved/unresolved agreement in 34/36 cells. Because this alternative resolves more rather than fewer cells, we use it only as a small-cluster sensitivity analysis and do not characterize it as more conservative than the primary hierarchical bootstrap.

We also perform a leave-one-split-out analysis. For each of the six omissions, the calculation is repeated using the remaining five split-level estimates and $t_{0.975,4}$. Among cells resolved in the primary analysis, 8/11 on $\Delta\mathrm{AUC}$, 5/8 on $-\Delta_{\mathrm{DP}}$, and 5/9 on $-\Delta_{\mathrm{EO}}$ remain resolved under every omission. Among primary unresolved cells, 21/25, 25/28, and 22/27, respectively, remain unresolved under every omission. Removing one split leaves only five top-level observations and widens the resulting interval, so we interpret this analysis as a finite-sample sensitivity check rather than as an alternative primary decision rule. All cells whose resolved label differs between the primary and six-split analyses have split-level sign consistency of 1.00, i.e., all six split-level estimates share the sign of the point estimate; in this analysis the split-level sign condition does not change any label beyond the interval and magnitude conditions. The accompanying artifact provides the cell-level and summary outputs for these analyses in `results/appendix_robustness/`.

D.5 Paired Uncertainty for Attribution Shifts

For paired configuration and protocol comparisons, we additionally estimate uncertainty directly for the attribution shift defined in Equation (7). Within each hierarchical-bootstrap replicate, the same resampled split–run units are retained across the two specifications, and we compute

$$\Delta^{(b)}_{\mathrm{attr}} = \widehat{\tau}^{(b)}_{-I\to+I}(P_b) - \widehat{\tau}^{(b)}_{-I\to+I}(P_a).$$

The reported interval is the 2.5th–97.5th percentile interval of these paired shift replicates. This distinguishes changes in the signs of two point estimates from direct uncertainty about whether attribution changes across specifications.

Table 17: **Controlled configuration-variation analyses.** "Pairs" denotes alternative-versus-primary configuration comparisons; each pair is evaluated on all three outcome coordinates. Published rows retain the hyperparameters associated with the corresponding released configuration, so variation-family labels are descriptive rather than factorial.

Method	Family	Alternative to primary	Datasets	Pairs
BIND	Budget	Deletion budget $1\% \to 10\%$ of the training set	Bail, Income	2
BeMap	Encoder	Primary GCN $\to$ GAT	Bail, Credit, Pokec-z	3
FairGNN	Upstream configuration	Released upstream GCN and GAT rows relative to the primary configuration	Pokec-z, Pokec-n	4
FairSIN	Encoder	GCN $\to$ GIN or SAGE	German, Bail, Credit, Pokec-z, Pokec-n	10
FairVGNN	Upstream configuration	Published GCN-spmm propagation row with its associated hyperparameters	German, Bail	2
FairVGNN	Encoder	Published GIN/SAGE rows with their associated hyperparameters	German, Bail, Credit where available	5
			Total	26

Table 18: **Published-horizon comparisons.** Method configuration, data, and validation-BCE checkpoint selection are held fixed within each pair; only the training horizon is set to its published value. Systematic and targeted comparison families are reported separately. The two SFG pairs whose published horizon equals the controlled horizon are reported as re-execution pairs in Table 13.

Family	Method	Configurations	Published-protocol change	Pairs
Systematic	FairSIN	5 datasets × GCN/GIN/SAGE	Published training horizon; validation-BCE selection retained	15
Systematic	SFG	Bail; SAGE	Published training horizon; validation-BCE selection retained	1
Targeted	FairGB	Bail, Credit; SAGE	$H = 1500$ and $H = 2000$, respectively	2
Targeted	FairVGNN	Bail; GCN	$H = 300$	1
Total				19

E Configuration-Variation Specifications

Configuration-variation analyses retain the controlled evaluation protocol and change a method-side specification relative to the corresponding primary configuration. Each alternative shares the method, dataset, data splits, run indices and seeds, checkpoint rule, evaluator, and external baseline with its primary counterpart. The resulting 26 alternative–primary pairs therefore assess within-method configuration sensitivity and are not treated as independent additions to the 36 primary cells.

For descriptive organization, we group the 26 pairs into 18 encoder variations, six upstream-configuration variations, and two intervention-budget variations. These labels identify the dominant specification difference rather than an isolated causal factor. In particular, published alternative rows can jointly change the backbone, propagation implementation, or hyperparameters associated with that row. We therefore interpret the resulting differences as configuration-level sensitivity of $\tau_{-I\to+I}(P)$, not as the contribution of a single hyperparameter or architectural component.

Within each pair, we compare the matched intervention contrast under the alternative and primary configurations,

$$\tau_{-I\to+I}(P; C_{\mathrm{alt}}) \quad \text{and} \quad \tau_{-I\to+I}(P; C_{\mathrm{primary}}),$$

while holding the controlled evaluation protocol P fixed. We summarize whether the point-estimate sign and resolved/unresolved reporting category change between the two configurations; the complete paired estimates are reported in Appendix H.

These 26 comparisons are a specification-sensitivity audit over the released alternatives that could be reconstructed under the controlled protocol. They are not intended to exhaust all plausible configurations for each method, and category-level counts should not be interpreted as prevalence estimates or as comparisons of which configuration family is intrinsically more sensitive.

F Protocol-Variation Specifications

Protocol variation holds the evaluated method configuration fixed while changing training, checkpoint selection, or other evaluation procedures. We distinguish checkpoint-selector replay, published-horizon comparisons, and broader published-procedure reconstructions because these families cover different cells and change different protocol factors. They are therefore reported separately rather than pooled into a common frequency or interpreted as nested levels of increasing protocol scope.

F.1 Checkpoint-Selector Replay

For each controlled trajectory, we replay checkpoint selection using two validation criteria. The primary selector minimizes validation binary cross-entropy, whereas the alternative selector maximizes validation AUC. No model is retrained and no additional training trajectory is generated; only the checkpoint selected from the stored trajectory changes. The comparison therefore measures selector sensitivity conditional on the realized training trajectory.

Selector replay is available for 65 controlled cells: the 36 primary cells, three task-adapted FnRGNN cells, and 26 configuration-variation cells. These roles remain separate in subsequent reporting. The

Table 19: **Published FairSIN training horizons used in the systematic comparisons.** Checkpoint selection remains validation BCE.

Dataset	GCN	GIN	SAGE
German	150	100	100
Bail	100	100	150
Credit	100	100	150
Pokec-z	100	100	150
Pokec-n	100	100	150

Table 20: **Targeted published-procedure comparisons.** Each row restores the listed method-specific procedure elements while retaining validation-BCE checkpoint selection. These comparisons are interpreted at the procedure level rather than as isolated changes in any single factor.

Method	Dataset	Published horizon	Additional published-procedure changes
NIFTY	German	1000	Training- and validation-view drop rates specified by the official command
FairGB	German	1500	Feature normalization disabled according to the upstream data-loading rule
FairVGNN	German	200	Feature normalization disabled according to the released dataset rule, including removal of harness and wrapper normalization
FairVGNN	Credit	200	The published Credit training loop, including the repository-default $\mathrm{clip}_c = 1.0$ setting

main protocol-sensitivity summary uses only the 36 primary cells; the additional replays are reported in Appendix I.1 and are not pooled with the primary selector counts.

F.2 Published-Horizon Comparisons

We restore a published training horizon only for methods whose released procedure does not use test information. Within each controlled–published pair, the method, dataset, method-side configuration, and validation-BCE checkpoint selection are held fixed; only the training horizon is replaced by its published counterpart. Because the selector is held fixed, these comparisons separate horizon changes from the selector changes analyzed in Appendix F.1. Table 18 summarizes the 19 resulting comparisons; the two SFG pairs whose published horizon equals the controlled horizon differ only by re-execution and are reported in Table 13.

For FairSIN, the published horizons are given in Table 19; at 100–150 epochs they are shorter than the controlled $H = 200$, so the systematic family mainly probes shorter training, and horizon extensions to 1000–2000 epochs occur only in the targeted families. For SFG, the published horizons are 200 epochs on German, 160 on Bail, and 200 on Credit.

The 16 systematic comparisons exhaust the admissible published-horizon configurations in the corresponding evaluated method families. The three targeted comparisons instead probe protocol sensitivity of specific frozen cells. Because the systematic and targeted families were constructed under different sampling rationales, they are not combined into a single frequency or success rate.

F.3 Targeted Published-Procedure Reconstructions

Four targeted comparisons require restoration of procedure-level factors beyond the training horizon. Table 18 lists these changes. Because several factors may change jointly, each comparison is interpreted as a contrast between the controlled and reconstructed published procedures as a whole rather than as an isolated horizon, preprocessing, or optimization contrast.

These four comparisons are targeted diagnostics of protocol sensitivity and are excluded from the systematic 16-cell published-horizon frequency. Their larger or smaller observed shifts cannot be attributed to protocol breadth alone because the evaluated cells and changed factors differ across comparison families.

F.4 Fixed Baseline Reference under Published Protocols

For every published-protocol evaluation, M^{-I} and M^{+I} follow the same published protocol. Their matched intervention contrast is therefore

$$\tau_{-I\to+I}(P_{\mathrm{pub}}) = \mathbb{E}_\omega \left[Y_\omega(M^{+I}; P_{\mathrm{pub}}) - Y_\omega(M^{-I}; P_{\mathrm{pub}}) \right],$$

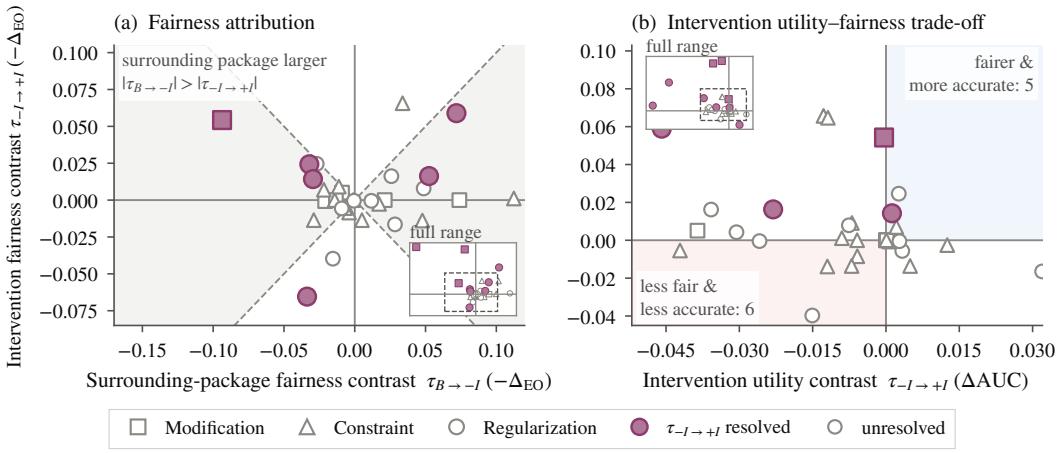


Figure 4: **Equal-opportunity counterpart to the primary attribution analysis.** Panel (a) compares the surrounding-package fairness contrast $\tau_{B\to -I}(P)$ with the matched intervention contrast $\tau_{-I\to +I}(P)$ on $-\Delta_{\mathrm{EO}}$. Dashed diagonals mark equal absolute magnitude, and the shaded region indicates cells for which the surrounding-package contrast is larger in absolute value. Panel (b) shows the matched intervention contrast in $\Delta\mathrm{AUC}$ versus $-\Delta_{\mathrm{EO}}$; positive values correspond to greater utility and lower equal-opportunity disparity, respectively. Marker shape denotes the descriptive method family, and filled markers denote resolved intervention contrasts on the plotted fairness coordinate. Insets show the full range where the main axes focus on the dense central region.

and remains a same-protocol intervention-level attribution target.

The external baseline B, however, remains the controlled $H = 200$ reference B_{200}. Consequently, contrasts involving B under a published method protocol are fixed-reference quantities:

$$\tau^{\text{fixed-ref}}_{B\to -I}(P_{\mathrm{pub}}) = \mathbb{E}_\omega\left[Y_\omega(M^{-I}; P_{\mathrm{pub}}) - Y_\omega(B_{200})\right],$$

and

$$\tau^{\text{fixed-ref}}_{B\to +I}(P_{\mathrm{pub}}) = \mathbb{E}_\omega\left[Y_\omega(M^{+I}; P_{\mathrm{pub}}) - Y_\omega(B_{200})\right].$$

These quantities must not be interpreted as same-protocol $B \to M$ comparisons under P_{pub}. The algebraic decomposition

$$\tau^{\text{fixed-ref}}_{B\to +I} = \tau^{\text{fixed-ref}}_{B\to -I} + \tau_{-I\to +I}(P_{\mathrm{pub}})$$

continues to hold because both baseline-involving contrasts use the same B_{200}. Accordingly, the controlled-versus-published protocol analyses in the main text compare only $\tau_{-I\to +I}$.

The accompanying artifact records the corresponding protocol family and the factors held fixed or varied for each reported comparison in `results/experiment_index.csv`.

G Primary Results: Cell-Level Estimates, Robustness, and Task-Adapted FnRGNN

This appendix reports the cell-level estimates underlying the primary package-versus-intervention analysis in Section 4.2.1. The 36 primary controlled cells are reported individually for $\Delta\mathrm{AUC}$, $-\Delta_{\mathrm{DP}}$, and $-\Delta_{\mathrm{EO}}$, followed by the separately analyzed task-adapted FnRGNN results. Aggregate counts in the main text are descriptive summaries of these included cells rather than estimates of prevalence across all fairness-aware GNN settings.

G.1 Primary Cell-Level Attribution Results

Figure 4 provides the equal-opportunity counterpart to the demographic-parity analysis in Figure 1. It uses the same attribution view in panel (a) and the same intervention-level utility–fairness view in panel (b), replacing $-\Delta_{\mathrm{DP}}$ with $-\Delta_{\mathrm{EO}}$.

Table 21: All 36 primary controlled cells on ΔAUC, grouped by method family as in Table 1. The task-adapted method FnRGNN is reported separately (Table 24).

Method	Dataset	$\tau_{B\to -I}$	95% CI	$\tau_{-I\to +I}$	95% CI	$\tau_{B\to +I}$
Modification ($n = 8$)						
BIND	bail [1pct]	$+0.105$	$[+0.057, +0.154]$	$+0.001$	$[-0.002, +0.004]$	$+0.106$
BIND	income [1pct]	$+0.127$	$[+0.117, +0.136]$	$+0.000$	$[-0.002, +0.001]$	$+0.126$
EDITS	bail	$+0.130$	$[+0.084, +0.178]$	-0.039	$[-0.099, +0.010]$	$+0.092$
EDITS	credit	$+0.011$	$[+0.005, +0.018]$	$\mathbf{-0.013}$	$[-0.020, -0.004]$	-0.002
EDITS	german	-0.011	$[-0.035, +0.016]$	-0.029	$[-0.064, +0.006]$	-0.040
FairEdit	bail	$+0.120$	$[+0.073, +0.167]$	$+0.000$	$[+0.000, +0.000]$	$+0.120$
FairEdit	credit	-0.019	$[-0.065, +0.008]$	$+0.000$	$[+0.000, +0.000]$	-0.019
FairEdit	german	-0.109	$[-0.179, -0.040]$	$+0.000$	$[+0.000, +0.000]$	-0.109
Constraint ($n = 13$)						
FairGNN	bail	$+0.127$	$[+0.080, +0.174]$	$+0.000$	$[+0.000, +0.000]$	$+0.127$
FairGNN	credit	-0.008	$[-0.020, +0.002]$	$+0.005$	$[-0.004, +0.014]$	-0.003
FairGNN	german	-0.061	$[-0.104, -0.022]$	-0.013	$[-0.041, +0.021]$	-0.074
FairGNN	pokec_n	$+0.000$	$[-0.002, +0.003]$	-0.006	$[-0.008, -0.004]$	-0.006
FairGNN	pokec_n_g	$+0.001$	$[-0.002, +0.003]$	-0.006	$[-0.008, -0.004]$	-0.005
FairGNN	pokec_z	$+0.024$	$[+0.018, +0.031]$	-0.007	$[-0.011, -0.003]$	$+0.017$
FairGNN	pokec_z_g	$+0.024$	$[+0.018, +0.031]$	-0.007	$[-0.010, -0.004]$	$+0.017$
FairVGNN	bail	$+0.114$	$[+0.066, +0.162]$	$\mathbf{-0.042}$	$[-0.062, -0.027]$	$+0.072$
FairVGNN	credit	$+0.013$	$[+0.006, +0.020]$	$\mathbf{-0.012}$	$[-0.016, -0.009]$	$+0.001$
FairVGNN	german	-0.043	$[-0.078, -0.009]$	$+0.013$	$[-0.017, +0.038]$	-0.030
SFG	bail [SAGE]	$+0.131$	$[+0.082, +0.179]$	$+0.002$	$[-0.008, +0.012]$	$+0.133$
SFG	credit [SAGE]	$+0.030$	$[+0.023, +0.037]$	-0.009	$[-0.012, -0.006]$	$+0.021$
SFG	german [SAGE]	$+0.005$	$[-0.019, +0.029]$	-0.012	$[-0.029, +0.004]$	-0.007
Regularization ($n = 15$)						
BeMap	bail	$+0.111$	$[+0.064, +0.159]$	$\mathbf{-0.139}$	$[-0.208, -0.091]$	-0.029
BeMap	credit	$+0.029$	$[+0.022, +0.036]$	$\mathbf{-0.036}$	$[-0.044, -0.031]$	-0.006
BeMap	pokec_z	$+0.053$	$[+0.047, +0.060]$	$\mathbf{-0.015}$	$[-0.024, -0.007]$	$+0.038$
FairGB	bail [SAGE]	$+0.122$	$[+0.074, +0.170]$	$\mathbf{+0.019}$	$[+0.007, +0.031]$	$+0.142$
FairGB	credit [SAGE]	$+0.031$	$[+0.025, +0.037]$	-0.008	$[-0.011, -0.004]$	$+0.023$
FairGB	german [SAGE]	-0.046	$[-0.082, -0.014]$	$\mathbf{+0.032}$	$[+0.010, +0.057]$	-0.014
FairSIN	bail	$+0.108$	$[+0.057, +0.158]$	$+0.003$	$[-0.004, +0.012]$	$+0.111$
FairSIN	credit	$+0.019$	$[+0.012, +0.026]$	$\mathbf{-0.046}$	$[-0.071, -0.031]$	-0.026
FairSIN	german	$+0.004$	$[-0.003, +0.011]$	$\mathbf{-0.110}$	$[-0.156, -0.060]$	-0.106
FairSIN	pokec_n	-0.012	$[-0.020, -0.004]$	$+0.003$	$[-0.002, +0.010]$	-0.009
FairSIN	pokec_z	$+0.025$	$[+0.019, +0.032]$	$+0.001$	$[-0.001, +0.004]$	$+0.026$
GEAR	bail [SAGE]	$+0.117$	$[+0.071, +0.163]$	$+0.003$	$[-0.001, +0.006]$	$+0.120$
NIFTY	bail	$+0.029$	$[-0.035, +0.092]$	-0.026	$[-0.050, -0.008]$	$+0.003$
NIFTY	credit	-0.122	$[-0.149, -0.097]$	$\mathbf{-0.023}$	$[-0.037, -0.013]$	-0.145
NIFTY	german	-0.153	$[-0.201, -0.106]$	-0.031	$[-0.056, -0.007]$	-0.183

Bold = resolved (unit-level sign consistency ≥ 0.75, $|\hat{\tau}| \geq 0.01$, and a 95% paired hierarchical-bootstrap interval excluding 0). All coordinates are oriented so that higher is better.

Tables 21–23 provide the cell-level values from which the aggregate primary summaries are obtained. In particular, they preserve the distinction among the surrounding-package contrast $\tau_{B\to -I}(P)$, the matched intervention contrast $\tau_{-I\to +I}(P)$, and the complete-method contrast $\tau_{B\to +I}(P)$. The intervention contrast is interpreted as a conditional contribution under the evaluated configuration and protocol; it does not by itself identify a mechanism.

G.2 Robustness of Primary Aggregate Summaries

The primary 36-cell summaries weight each supported method–dataset cell equally. Because coverage is non-Cartesian, we assess whether the headline package-versus-intervention magnitude comparison is driven by methods with broader dataset coverage or by intervention boundaries that jointly activate multiple fairness-specific components. We therefore consider three descriptive robustness summaries.

First, we repeat the analysis on a conservative single-operation intervention subset, defined from the intervention constructions in Table 5 without reference to the observed contrasts. This subset contains BIND, BeMap, FairEdit, GEAR, and NIFTY (12 cells), for which M^{-I} and M^{+I} differ by one designated fairness operation or objective rather than by a coupled set of fairness-specific

Table 22: All 36 primary controlled cells on $-\Delta_{\mathrm{DP}}$, grouped by method family as in Table 1. The task-adapted method FnRGNN is reported separately (Table 24).

Method	Dataset	$\tau_{B\to -I}$	95% CI	$\tau_{-I\to +I}$	95% CI	$\tau_{B\to +I}$
Modification ($n = 8$)						
BIND	bail [1pct]	-0.041	$[-0.057, -0.023]$	$+0.001$	$[-0.002, +0.003]$	-0.040
BIND	income [1pct]	-0.065	$[-0.091, -0.030]$	$\mathbf{+0.042}$	$[+0.036, +0.049]$	-0.023
EDITS	bail	-0.036	$[-0.046, -0.025]$	$+0.007$	$[-0.003, +0.020]$	-0.028
EDITS	credit	-0.266	$[-0.299, -0.234]$	$\mathbf{+0.209}$	$[+0.076, +0.340]$	-0.057
EDITS	german	-0.041	$[-0.122, +0.032]$	$\mathbf{+0.276}$	$[+0.118, +0.399]$	$+0.235$
FairEdit	bail	-0.040	$[-0.055, -0.025]$	$+0.000$	$[+0.000, +0.000]$	-0.040
FairEdit	credit	$+0.077$	$[+0.048, +0.107]$	$+0.000$	$[+0.000, +0.000]$	$+0.077$
FairEdit	german	$+0.065$	$[-0.085, +0.210]$	$+0.001$	$[-0.002, +0.002]$	$+0.066$
Constraint ($n = 13$)						
FairGNN	bail	-0.039	$[-0.050, -0.028]$	$+0.000$	$[+0.000, +0.000]$	-0.039
FairGNN	credit	$+0.001$	$[-0.031, +0.033]$	-0.010	$[-0.027, +0.005]$	-0.009
FairGNN	german	$+0.057$	$[-0.006, +0.124]$	$+0.093$	$[-0.002, +0.185]$	$+0.150$
FairGNN	pokec_n	-0.004	$[-0.012, +0.007]$	-0.004	$[-0.016, +0.007]$	-0.008
FairGNN	pokec_n_g	-0.004	$[-0.011, +0.003]$	-0.005	$[-0.011, +0.000]$	-0.009
FairGNN	pokec_z	-0.036	$[-0.058, -0.016]$	-0.018	$[-0.034, -0.005]$	-0.054
FairGNN	pokec_z_g	-0.002	$[-0.012, +0.008]$	$+0.001$	$[-0.006, +0.009]$	-0.001
FairVGNN	bail	-0.030	$[-0.041, -0.019]$	$+0.007$	$[-0.003, +0.017]$	-0.023
FairVGNN	credit	$+0.039$	$[+0.014, +0.061]$	$+0.000$	$[-0.041, +0.033]$	$+0.039$
FairVGNN	german	$+0.046$	$[-0.059, +0.167]$	-0.003	$[-0.110, +0.103]$	$+0.043$
SFG	bail [SAGE]	-0.034	$[-0.058, -0.011]$	$+0.020$	$[+0.000, +0.045]$	-0.014
SFG	credit [SAGE]	$+0.114$	$[+0.086, +0.143]$	-0.001	$[-0.023, +0.015]$	$+0.113$
SFG	german [SAGE]	$+0.141$	$[+0.015, +0.265]$	$+0.073$	$[+0.020, +0.127]$	$+0.215$
Regularization ($n = 15$)						
BeMap	bail	-0.058	$[-0.072, -0.044]$	$\mathbf{+0.060}$	$[+0.042, +0.074]$	$+0.002$
BeMap	credit	$+0.029$	$[-0.002, +0.057]$	$+0.022$	$[-0.020, +0.067]$	$+0.052$
BeMap	pokec_z	-0.024	$[-0.055, +0.005]$	-0.067	$[-0.144, +0.007]$	-0.091
FairGB	bail [SAGE]	-0.031	$[-0.071, +0.004]$	-0.060	$[-0.108, -0.016]$	-0.091
FairGB	credit [SAGE]	$+0.049$	$[+0.017, +0.081]$	$+0.014$	$[-0.018, +0.050]$	$+0.063$
FairGB	german [SAGE]	$+0.054$	$[-0.076, +0.175]$	-0.027	$[-0.095, +0.032]$	$+0.026$
FairSIN	bail	-0.038	$[-0.051, -0.025]$	$+0.005$	$[+0.001, +0.010]$	-0.033
FairSIN	credit	$+0.073$	$[+0.047, +0.098]$	$\mathbf{+0.077}$	$[+0.055, +0.099]$	$+0.150$
FairSIN	german	$+0.133$	$[+0.051, +0.239]$	$\mathbf{+0.179}$	$[+0.083, +0.276]$	$+0.312$
FairSIN	pokec_n	$+0.011$	$[-0.007, +0.032]$	-0.002	$[-0.018, +0.011]$	$+0.009$
FairSIN	pokec_z	-0.034	$[-0.050, -0.017]$	$\mathbf{+0.018}$	$[+0.010, +0.028]$	-0.016
GEAR	bail [SAGE]	-0.032	$[-0.068, +0.002]$	$+0.016$	$[-0.020, +0.050]$	-0.016
NIFTY	bail	-0.002	$[-0.013, +0.008]$	$+0.001$	$[-0.006, +0.009]$	-0.001
NIFTY	credit	$+0.078$	$[+0.044, +0.113]$	$\mathbf{+0.017}$	$[+0.006, +0.028]$	$+0.095$
NIFTY	german	$+0.238$	$[+0.110, +0.369]$	-0.006	$[-0.042, +0.032]$	$+0.232$

Bold = resolved (unit-level sign consistency ≥ 0.75, $|\hat{\tau}| \geq 0.01$, and a 95% paired hierarchical-bootstrap interval excluding 0). All coordinates are oriented so that higher is better.

components. Second, we compute the package-larger proportion within each method and average these proportions equally across methods. Third, we recompute the original cell-weighted summary after leaving out each method in turn. We further exclude the three FairEdit cells from the single-operation subset, because the released FairEdit configuration deletes only ten edges, and separately exclude the two cells whose intervention effect falls below a re-execution threshold fixed before the re-execution results were examined (FairEdit/Bail and FairEdit/Credit; Appendix C). Finally, we repeat the cell-weighted, method-balanced, and leave-one-method-out summaries with the external baseline trained for 1000 epochs on every dataset.

For FairGNN/Bail, M^{-I} and M^{+I} differ in all 30 matched units, so the configured α and β do alter training, but the resulting change in test outcomes is small (at most 8×10^{-4} in unit-level AUC); this cell satisfies the exclusion rule on two of three coordinates and is therefore retained.

Under the reported external baseline, the surrounding-package contrast is larger than the matched intervention contrast in a majority of the audited settings on every coordinate and in every summary, including the single-operation subset without FairEdit and the analysis excluding inactive interventions; method-balanced proportions exceed cell-weighted ones, indicating that methods with more cells have lower package-larger proportions. With the more strongly trained reference, the proportions fall to about one half, and the leave-one-method-out range for $-\Delta_{\mathrm{DP}}$ extends below one

Table 23: All 36 primary controlled cells on $-\Delta_{\mathrm{EO}}$, grouped by method family as in Table 1. The task-adapted method FnRGNN is reported separately (Table 24).

Method	Dataset	$\tau_{B\to -I}$	95% CI	$\tau_{-I\to +I}$	95% CI	$\tau_{B\to +I}$
Modification ($n = 8$)						
BIND	bail [1pct]	-0.021	$[-0.038, -0.003]$	-0.001	$[-0.007, +0.006]$	-0.022
BIND	income [1pct]	-0.094	$[-0.136, -0.050]$	$\mathbf{+0.054}$	$[+0.036, +0.070]$	-0.039
EDITS	bail	-0.009	$[-0.019, +0.000]$	$+0.005$	$[-0.001, +0.012]$	-0.004
EDITS	credit	-0.327	$[-0.362, -0.292]$	$\mathbf{+0.233}$	$[+0.097, +0.363]$	-0.094
EDITS	german	-0.060	$[-0.141, +0.020]$	$\mathbf{+0.221}$	$[+0.087, +0.338]$	$+0.161$
FairEdit	bail	-0.013	$[-0.023, -0.003]$	$+0.000$	$[+0.000, +0.000]$	-0.013
FairEdit	credit	$+0.074$	$[+0.045, +0.104]$	$+0.000$	$[+0.000, +0.000]$	$+0.074$
FairEdit	german	$+0.021$	$[-0.128, +0.168]$	$+0.000$	$[-0.003, +0.002]$	$+0.021$
Constraint ($n = 13$)						
FairGNN	bail	-0.014	$[-0.023, -0.006]$	$+0.000$	$[+0.000, +0.000]$	-0.014
FairGNN	credit	$+0.005$	$[-0.029, +0.037]$	-0.014	$[-0.033, +0.004]$	-0.009
FairGNN	german	$+0.034$	$[-0.033, +0.099]$	$+0.066$	$[-0.023, +0.148]$	$+0.099$
FairGNN	pokec_n	-0.001	$[-0.010, +0.007]$	$+0.000$	$[-0.011, +0.013]$	-0.001
FairGNN	pokec_n_g	-0.004	$[-0.012, +0.003]$	-0.008	$[-0.015, -0.002]$	-0.012
FairGNN	pokec_z	-0.029	$[-0.048, -0.009]$	-0.014	$[-0.026, -0.002]$	-0.043
FairGNN	pokec_z_g	-0.011	$[-0.023, +0.000]$	$+0.009$	$[-0.001, +0.019]$	-0.002
FairVGNN	bail	-0.007	$[-0.019, +0.005]$	-0.005	$[-0.012, +0.001]$	-0.012
FairVGNN	credit	$+0.048$	$[+0.021, +0.072]$	-0.014	$[-0.060, +0.021]$	$+0.034$
FairVGNN	german	$+0.017$	$[-0.093, +0.149]$	-0.003	$[-0.117, +0.114]$	$+0.015$
SFG	bail [SAGE]	-0.022	$[-0.053, +0.003]$	$+0.007$	$[-0.007, +0.027]$	-0.015
SFG	credit [SAGE]	$+0.112$	$[+0.086, +0.140]$	$+0.001$	$[-0.012, +0.013]$	$+0.113$
SFG	german [SAGE]	$+0.126$	$[+0.006, +0.242]$	$+0.065$	$[+0.018, +0.115]$	$+0.191$
Regularization ($n = 15$)						
BeMap	bail	-0.032	$[-0.048, -0.016]$	$\mathbf{+0.024}$	$[+0.016, +0.036]$	-0.008
BeMap	credit	$+0.026$	$[-0.008, +0.056]$	$+0.016$	$[-0.029, +0.063]$	$+0.042$
BeMap	pokec_z	-0.015	$[-0.044, +0.010]$	-0.040	$[-0.097, +0.014]$	-0.055
FairGB	bail [SAGE]	-0.034	$[-0.076, -0.004]$	$\mathbf{-0.065}$	$[-0.112, -0.023]$	-0.099
FairGB	credit [SAGE]	$+0.049$	$[+0.016, +0.081]$	$+0.008$	$[-0.022, +0.042]$	$+0.057$
FairGB	german [SAGE]	$+0.028$	$[-0.096, +0.141]$	-0.016	$[-0.075, +0.034]$	$+0.012$
FairSIN	bail	-0.009	$[-0.025, +0.006]$	-0.006	$[-0.014, +0.001]$	-0.015
FairSIN	credit	$+0.072$	$[+0.044, +0.099]$	$\mathbf{+0.059}$	$[+0.037, +0.083]$	$+0.131$
FairSIN	german	$+0.128$	$[+0.049, +0.226]$	$\mathbf{+0.132}$	$[+0.054, +0.210]$	$+0.261$
FairSIN	pokec_n	$+0.012$	$[-0.006, +0.032]$	-0.001	$[-0.013, +0.011]$	$+0.011$
FairSIN	pokec_z	-0.030	$[-0.048, -0.011]$	$\mathbf{+0.014}$	$[+0.008, +0.021]$	-0.015
GEAR	bail [SAGE]	-0.027	$[-0.057, -0.003]$	$+0.025$	$[+0.000, +0.048]$	-0.002
NIFTY	bail	-0.001	$[-0.008, +0.006]$	$+0.000$	$[-0.005, +0.004]$	-0.001
NIFTY	credit	$+0.052$	$[+0.015, +0.091]$	$\mathbf{+0.016}$	$[+0.005, +0.029]$	$+0.069$
NIFTY	german	$+0.188$	$[+0.062, +0.317]$	$+0.004$	$[-0.034, +0.044]$	$+0.192$

Bold = resolved (unit-level sign consistency ≥ 0.75, $|\hat{\tau}| \geq 0.01$, and a 95% paired hierarchical-bootstrap interval excluding 0). All coordinates are oriented so that higher is better.

half. The magnitude comparison therefore depends on how strongly the external baseline is trained, whereas the matched intervention contrasts themselves do not depend on B.

G.3 Task-Adapted FnRGNN Results

FnRGNN is reported separately because its released implementation targets node regression and therefore requires task-level adaptation for the common node-classification evaluation. The matched construction and shared task adaptations are specified in Appendix A.3. The three classification cells cover German, Pokec-z, and Pokec-n and are not included in any aggregate over the 36 primary controlled cells.

For the adapted node-classification task, the matched intervention contribution is positive and resolved on both fairness coordinates for all three evaluated datasets, while the $\Delta\mathrm{AUC}$ contribution is negative and resolved on Pokec-z and Pokec-n. Table 24 additionally reports the released node-regression task as a task-specific reference. Because classification and regression use different predictive outcomes and metric scales, we do not compare intervention magnitudes across the two tasks; the regression results are included only to show how the same method-specific intervention is represented under its released task.

Table 24: FnRGNN, a method released for node regression, evaluated on the common classification task (loss switched to BCE) and on its own regression task (released MSE loss). Same split, released configuration and harness-completed training loop. Reported on its own and never pooled with Table 1; regression coordinates are in standardised target units and are comparable only with each other.

Coordinate	Dataset	$\tau_{B\to -I}$	95% CI	$\tau_{-I\to +I}$	95% CI	$\tau_{B\to +I}$	95% CI
Classification (BCE, validation-BCE selector)							
ΔAUC	german	-0.032	$[-0.049, -0.017]$	-0.006	$[-0.038, +0.024]$	-0.038	$[-0.074, -0.006]$
	pokec_n	$+0.023$	$[+0.015, +0.030]$	$\mathbf{-0.020}$	$[-0.031, -0.007]$	$+0.003$	$[-0.004, +0.011]$
	pokec_z	$+0.037$	$[+0.028, +0.047]$	$\mathbf{-0.062}$	$[-0.087, -0.040]$	-0.025	$[-0.045, -0.006]$
$-\Delta_{\mathrm{DP}}$	german	$+0.015$	$[-0.062, +0.105]$	$\mathbf{+0.160}$	$[+0.019, +0.300]$	$+0.175$	$[+0.015, +0.329]$
	pokec_n	-0.003	$[-0.018, +0.015]$	$\mathbf{+0.066}$	$[+0.020, +0.114]$	$+0.063$	$[+0.011, +0.121]$
	pokec_z	-0.028	$[-0.055, +0.000]$	$\mathbf{+0.080}$	$[+0.047, +0.114]$	$+0.052$	$[+0.014, +0.091]$
$-\Delta_{\mathrm{EO}}$	german	-0.024	$[-0.102, +0.064]$	$\mathbf{+0.158}$	$[+0.018, +0.296]$	$+0.134$	$[-0.019, +0.280]$
	pokec_n	-0.006	$[-0.025, +0.015]$	$\mathbf{+0.063}$	$[+0.018, +0.110]$	$+0.057$	$[+0.003, +0.117]$
	pokec_z	-0.020	$[-0.047, +0.010]$	$\mathbf{+0.060}$	$[+0.035, +0.085]$	$+0.040$	$[+0.007, +0.074]$
Regression (released task, MSE, validation-MSE selector; standardised units)							
$-$MSE	german	$+0.734$	$[+0.454, +1.047]$	$\mathbf{-0.112}$	$[-0.152, -0.080]$	$+0.622$	$[+0.352, +0.932]$
	pokec_n	$+0.082$	$[+0.067, +0.097]$	$\mathbf{+0.350}$	$[+0.331, +0.369]$	$+0.432$	$[+0.405, +0.457]$
	pokec_z	$+0.095$	$[+0.079, +0.113]$	$\mathbf{+0.394}$	$[+0.374, +0.418]$	$+0.490$	$[+0.460, +0.521]$
$-$MeanGap	german	-0.021	$[-0.077, +0.034]$	$\mathbf{+0.143}$	$[+0.105, +0.183]$	$+0.122$	$[+0.054, +0.185]$
	pokec_n	-0.010	$[-0.040, +0.017]$	$\mathbf{+0.033}$	$[+0.012, +0.056]$	$+0.023$	$[-0.007, +0.057]$
	pokec_z	-0.044	$[-0.099, +0.007]$	$\mathbf{+0.053}$	$[+0.004, +0.107]$	$+0.009$	$[-0.013, +0.027]$
$-$WD	german	-0.020	$[-0.080, +0.039]$	$\mathbf{+0.122}$	$[+0.100, +0.141]$	$+0.102$	$[+0.041, +0.162]$
	pokec_n	-0.029	$[-0.065, +0.005]$	$\mathbf{+0.036}$	$[+0.016, +0.060]$	$+0.007$	$[-0.015, +0.031]$
	pokec_z	-0.054	$[-0.096, -0.014]$	$\mathbf{+0.051}$	$[+0.018, +0.090]$	-0.003	$[-0.022, +0.013]$

Bold = resolved (unit-level sign consistency ≥ 0.75, $|\hat{\tau}| \geq 0.01$, and a 95% paired hierarchical-bootstrap interval excluding 0). All coordinates are oriented so that higher is better.

Of the three FnRGNN components, MMD alignment acts on hidden representations, whereas edge reweighting acts on the input graph and GWN normalisation on the prediction layer; the classification adaptation therefore changes the output space in which GWN operates, which is one reason the task-adapted cells are never pooled with the primary evaluation.

H Additional Configuration-Sensitivity Results

Table 25 reports the complete paired results for the 26 configuration variations specified in Appendix E. Each row compares the matched intervention contrast under an alternative method configuration with that under its corresponding primary configuration while retaining the controlled evaluation protocol.

Bold values denote resolved contrasts. A dagger (†) denotes a point-estimate sign change between the primary and alternative configurations, and a double dagger (‡) denotes a change in resolved status. All indicators are computed from the unrounded estimates. A resolved-status change may therefore represent either a resolved-to-unresolved or an unresolved-to-resolved transition and is not interpreted as a directional improvement or deterioration. Because the configuration families contain unequal numbers of comparisons and the alternatives are not a factorial design, the table is used to assess cell-level specification sensitivity rather than to rank encoder, upstream-configuration, or intervention-budget variations by sensitivity.

I Additional Protocol-Sensitivity Results

This appendix reports the complete paired estimates underlying the protocol-sensitivity analysis in Section 4.2.2. The tables report the matched intervention contrast $\tau_{-I\to +I}(P)$ on $-\Delta_{\mathrm{DP}}$, corresponding to Figure 2. The three comparison families are reported separately because they cover different cells and change different protocol factors. Bold values denote resolved contrasts, a dagger (†) denotes a point-estimate sign change between protocols, and a double dagger (‡) denotes a change in resolved status. All indicators are computed from the unrounded estimates.

Table 25: **Paired configuration-sensitivity results.** Entries report $\tau_{-I\to+I}(P)$ as primary → alternative. Bold values are resolved. † denotes a point-estimate sign flip and ‡ a change in resolved status. § marks a coordinate whose paired Δ_{attr} 95% interval excludes zero. Symbols are determined from the unrounded estimates.

Method	Data	Primary → alternative	ΔAUC	$-\Delta_{\text{DP}}$	$-\Delta_{\text{EO}}$
Encoder variations ($n = 18$)					
BeMap	Bail	GCN → GAT	**-0.139** → **-0.097**§	**+0.060** → **+0.062**	**+0.024** → **+0.037**
BeMap	Credit	GCN → GAT	**-0.036** → -0.009‡§	+0.022 → +0.029	+0.016 → +0.022
BeMap	Pokec-z	GCN → GAT	**-0.015** → -0.005‡§	-0.067 → **-0.038**‡	-0.040 → -0.019
FairSIN	Bail	GCN → GIN	+0.003 → +0.007	+0.005 → +0.000	-0.006 → -0.011
FairSIN	Bail	GCN → SAGE	+0.003 → +0.013	+0.005 → -0.002†	-0.006 → -0.005
FairSIN	Credit	GCN → GIN	**-0.046** → **-0.035**	**+0.077** → **+0.060**	**+0.059** → **+0.036**§
FairSIN	Credit	GCN → SAGE	**-0.046** → **-0.014**§	**+0.077** → -0.049†‡§	**+0.059** → -0.038†‡§
FairSIN	German	GCN → GIN	**-0.110** → -0.003‡§	**+0.179** → +0.001‡§	**+0.132** → -0.009†‡§
FairSIN	German	GCN → SAGE	**-0.110** → -0.015‡§	**+0.179** → +0.005‡§	**+0.132** → +0.014‡§
FairSIN	Pokec-n	GCN → GIN	+0.003 → +0.005	-0.002 → +0.007†	-0.001 → +0.001†
FairSIN	Pokec-n	GCN → SAGE	+0.003 → -0.000†	-0.002 → +0.002†	-0.001 → +0.010†
FairSIN	Pokec-z	GCN → GIN	+0.001 → +0.005	**+0.018** → +0.004‡	**+0.014** → +0.005‡
FairSIN	Pokec-z	GCN → SAGE	+0.001 → -0.003†	**+0.018** → -0.003†‡§	**+0.014** → -0.004†‡§
FairVGNN	Bail	GCN → SAGE	**-0.042** → +0.012†‡§	+0.007 → +0.024	-0.005 → +0.019†
FairVGNN	Credit	GCN → GIN	**-0.012** → **-0.072**§	-0.000 → **+0.096**†‡§	-0.014 → **+0.080**†‡§
FairVGNN	Credit	GCN → SAGE	**-0.012** → -0.009‡	-0.000 → +0.023†	-0.014 → +0.014†
FairVGNN	German	GCN → GIN	+0.013 → +0.021	-0.003 → **+0.081**†‡§	-0.003 → **+0.088**†‡§
FairVGNN	German	GCN → SAGE	+0.013 → +0.024	-0.003 → +0.033†	-0.003 → +0.039†
Upstream-configuration variations ($n = 6$)					
FairGNN	Pokec-n	default → upstream GAT	-0.006 → +0.005†	-0.004 → +0.048†	+0.000 → +0.040
FairGNN	Pokec-n	default → upstream GCN	-0.006 → **-0.039**‡§	-0.004 → +0.026†	+0.000 → +0.027
FairGNN	Pokec-z	default → upstream GAT	-0.007 → -0.045§	-0.018 → -0.015	-0.014 → -0.026
FairGNN	Pokec-z	default → upstream GCN	-0.007 → **-0.045**‡§	-0.018 → +0.028†§	-0.014 → +0.030†§
FairVGNN	Bail	default → GCN-spmm	**-0.042** → **-0.041**	+0.007 → +0.008	-0.005 → -0.003
FairVGNN	German	default → GCN-spmm	+0.013 → **+0.023**‡	-0.003 → +0.046†	-0.003 → +0.064†
Intervention-budget variations ($n = 2$)					
BIND	Bail	1% → 10%	+0.001 → -0.001†	+0.001 → +0.003	-0.001 → -0.000
BIND	Income	1% → 10%	-0.000 → -0.000	**+0.042** → **+0.053**§	**+0.054** → **+0.070**§

Notes. Each entry shows the primary and alternative point estimates of $\tau_{-I\to+I}(P)$. Bold values satisfy the resolved criterion. † denotes a change in the sign of the unrounded point estimate; ‡ denotes a change in resolved status; § denotes that the paired shift Δ_{attr} has a 95% interval excluding zero, estimated directly on the 30 matched units rather than read off the two intervals. Sign flips near zero should therefore not be interpreted independently of the resolution status. All comparisons retain the controlled protocol and are paired with their corresponding primary configuration.

I.1 Checkpoint-Selector Sensitivity

Table 26 reports the 36 primary controlled cells used in Figure 2(a). Both estimates use the same stored training trajectory; only checkpoint selection changes from validation BCE to validation AUC. The resulting comparison therefore measures selector sensitivity conditional on the realized trajectory rather than sensitivity to retraining.

The selector replay is also available for three task-adapted FnRGNN cells and 26 configuration-variation cells. These 29 additional replays are excluded from the 36-cell primary selector summary because they belong to different evaluation roles. Their machine-readable paired estimates are provided in the accompanying artifact in `results/3a_protocol_selector_bce_vs_auc.csv`.

I.2 Published Horizon

Table 27 reports the 19 published-horizon comparisons shown in Figure 2(b). In every row, the method configuration, data, and validation-BCE checkpoint selection are held fixed while the training horizon is changed from the controlled specification to its published value. The 16 systematic and three targeted comparisons are shown in the same table for compactness but remain separate evaluation families and are not pooled into a single frequency.

Table 26: **Checkpoint-selector sensitivity for the 36 primary cells.** Entries report $\tau_{-I\to+I}(P)$ on $-\Delta_{\mathrm{DP}}$ as validation BCE $\to$ validation AUC. Bold values are resolved. † denotes a point-estimate sign change and ‡ a change in resolved status. Symbols are computed from unrounded estimates.

Method	Dataset	BCE $\to$ AUC
BIND	Bail	$+0.001 \to -0.002^{\dagger}$
BIND	Income	$\mathbf{+0.042} \to +0.001^{\ddagger}$
BeMap	Bail	$\mathbf{+0.060} \to \mathbf{+0.044}$
BeMap	Credit	$+0.022 \to \mathbf{+0.042}^{\ddagger}$
BeMap	Pokec-z	$-0.067 \to -0.069$
EDITS	Bail	$+0.007 \to +0.010$
EDITS	Credit	$\mathbf{+0.209} \to \mathbf{+0.174}$
EDITS	German	$\mathbf{+0.276} \to \mathbf{+0.287}$
FairEdit	Bail	$+{<}0.001 \to -{<}0.001^{\dagger}$
FairEdit	Credit	$+0.000 \to +0.000$
FairEdit	German	$+0.001 \to +{<}0.001$
FairGB	Bail	$-0.060 \to \mathbf{-0.081}^{\ddagger}$
FairGB	Credit	$+0.014 \to -0.016^{\dagger}$
FairGB	German	$-0.027 \to \mathbf{-0.077}^{\ddagger}$
FairGNN	Bail	$-{<}0.001 \to -{<}0.001$
FairGNN	Credit	$-0.010 \to +0.004^{\dagger}$
FairGNN	German	$+0.093 \to +0.126$
FairGNN	Pokec-n	$-0.004 \to +0.003^{\dagger}$
FairGNN	Pokec-n-g	$-0.005 \to -0.004$
FairGNN	Pokec-z	$-0.018 \to -0.008$
FairGNN	Pokec-z-g	$+0.001 \to -0.005^{\dagger}$
FairSIN	Bail	$+0.005 \to +0.010$
FairSIN	Credit	$\mathbf{+0.077} \to \mathbf{+0.117}$
FairSIN	German	$\mathbf{+0.179} \to \mathbf{+0.178}$
FairSIN	Pokec-n	$-0.002 \to \mathbf{+0.017}^{\dagger\ddagger}$
FairSIN	Pokec-z	$\mathbf{+0.018} \to \mathbf{+0.017}$
FairVGNN	Bail	$+0.007 \to +{<}0.001$
FairVGNN	Credit	$-{<}0.001 \to -0.032$
FairVGNN	German	$-0.003 \to -0.063$
GEAR	Bail	$+0.016 \to +0.020$
NIFTY	Bail	$+0.001 \to +0.001$
NIFTY	Credit	$\mathbf{+0.017} \to \mathbf{+0.014}$
NIFTY	German	$-0.006 \to \mathbf{+0.024}^{\dagger\ddagger}$
SFG	Bail	$+0.020 \to -0.001^{\dagger}$
SFG	Credit	$-0.001 \to +0.010^{\dagger}$
SFG	German	$+0.073 \to +0.013$

Notes. The two estimates use the same stored training trajectory. Only the validation checkpoint-selection criterion changes. Point-estimate sign changes close to zero should be interpreted together with the resolved status.

Table 27: **Controlled versus published training horizons.** Entries report $\tau_{-I\to+I}(P)$ on $-\Delta_{\mathrm{DP}}$ as controlled $\to$ published horizon. Systematic and targeted comparisons are shown separately. Bold values are resolved; † denotes a sign change and ‡ a resolution change.

Method	Dataset	Configuration	Controlled $\to$ published
Systematic evaluations ($n = 16$)			
FairSIN	Bail	GCN	$+0.005 \to +0.005$
FairSIN	Bail	GIN	$+{<}0.001 \to +0.003$
FairSIN	Bail	SAGE	$-0.002 \to -0.003$
FairSIN	Credit	GCN	$\mathbf{+0.077} \to \mathbf{+0.083}$
FairSIN	Credit	GIN	$\mathbf{+0.060} \to \mathbf{+0.054}$
FairSIN	Credit	SAGE	$-0.049 \to -0.025$
FairSIN	German	GCN	$\mathbf{+0.179} \to \mathbf{+0.149}$
FairSIN	German	GIN	$+0.001 \to +0.035$
FairSIN	German	SAGE	$+0.005 \to +0.016$
FairSIN	Pokec-n	GCN	$-0.002 \to -0.001$
FairSIN	Pokec-n	GIN	$+0.007 \to +0.015$
FairSIN	Pokec-n	SAGE	$+0.002 \to +0.002$
FairSIN	Pokec-z	GCN	$\mathbf{+0.018} \to \mathbf{+0.016}$
FairSIN	Pokec-z	GIN	$+0.004 \to -0.009^{\dagger}$
FairSIN	Pokec-z	SAGE	$-0.003 \to -0.003$
SFG	Bail	Default	$+0.020 \to +0.016$
Targeted validations ($n = 3$)			
FairGB	Bail	Default	$-0.060 \to -0.010$
FairGB	Credit	Default	$+0.014 \to \mathbf{+0.081}^{\ddagger}$
FairVGNN	Bail	Default	$+0.007 \to +0.003$
Re-execution only ($n = 2$)			
SFG	Credit	Default	$-0.001 \to -0.001$
SFG	German	Default	$+0.073 \to +0.028$

Notes. Only the training horizon differs between the two protocols; checkpoint selection is validation BCE in both. The systematic and targeted blocks have different sampling rationales and are not pooled into a common success rate. SFG on German and Credit is listed separately: its two protocols resolve to the same configuration and the same 200-epoch horizon, so that comparison measures re-execution rather than the published horizon, and it is excluded from the systematic count.

Table 28: **Targeted published-procedure comparisons.** Entries report $\tau_{-I\to+I}(P)$ on $-\Delta_{\mathrm{DP}}$ as controlled $\to$ published procedure. Validation-BCE checkpoint selection is retained in every row. Bold values are resolved; † denotes a sign change and ‡ a resolution change.

Method	Dataset	Controlled $\to$ published	Factors changed
FairGB	German	$-0.027 \to +0.020^{\dagger}$	$H: 200 \to 1500$; feature normalization disabled
FairVGNN	Credit	$-{<}0.001 \to +\mathbf{0.032}^{\dagger\ddagger}$	Published Credit training loop ($\mathrm{clip}_c = 1.0$); horizon unchanged
FairVGNN	German	$-0.003 \to +\mathbf{0.190}^{\dagger\ddagger}$	Feature normalization disabled in both the harness and wrapper; horizon unchanged
NIFTY	German	$-0.006 \to -\mathbf{0.151}^{\ddagger}$	$H: 200 \to 1000$; published training- and validation-view drop rates

Notes. Multiple procedure components change jointly in every row. These comparisons therefore quantify shifts under the published procedure as a bundle and do not identify the effect of any individual factor. All four change either sign or resolved status; three change sign and three change resolved status.

I.3 Published-Procedure Comparisons

Table 28 reports the four targeted comparisons shown in Figure 2(c). These reconstructions restore method-specific procedure elements beyond the training horizon, while retaining validation-BCE checkpoint selection. Because several factors change jointly, the resulting differences are interpreted as procedure-level sensitivity and are not attributed to any one protocol component in isolation.

All tables in this appendix concern $\tau_{-I\to+I}$, for which M^{-I} and M^{+I} are evaluated under the same protocol within each row. The fixed controlled baseline used for contrasts involving B and the resulting interpretation of those fixed-reference quantities are specified separately in Appendix F.4.

J Mechanistic Case-Study Details

The analyses in this appendix are targeted diagnostic follow-ups rather than samples intended to represent the benchmark as a whole. Case selection occurred after the protocol-sensitivity audit, using large or diagnostically informative attribution shifts as triggers. The subsequent follow-up design, analysis sequence, and stopping rules were frozen before the corresponding diagnostic reruns. Thus, "pre-specified" in this appendix refers to the follow-up procedure rather than to selection of the cases before observing the protocol-sensitivity results. The accompanying artifact retains the corresponding frozen pre-run plans and stop reports.

J.1 Case Selection, Analysis Order, and Stopping Rule

Table 29 summarizes the three planned follow-up cases. The screening values are the frozen protocol-comparison estimates used to motivate the follow-up; dedicated trajectory reruns reported later produce their own matched estimates.

NIFTY/German requires an additional factorial step because its reconstructed published procedure changes both the training horizon H and the training/validation-view drop-rate setting D. FairGB/Bail and FairGB/Credit were paired as cross-method follow-ups within FairGB because their screening contrasts move in different directions. FairGB/Credit did not satisfy the frozen instrumentation-invariance criterion, and the analysis plan did not permit adding repetitions to pass the gate; no mechanistic decomposition is therefore reported for that cell.

J.2 NIFTY/German Factorial Protocol Decomposition

The published NIFTY/German reconstruction changes two protocol factors relative to the controlled evaluation: the training horizon H and the training/validation-view drop-rate setting D. We evaluate

$$(H_0, D_0), \qquad (H_1, D_0), \qquad (H_0, D_1), \qquad (H_1, D_1),$$

where $H_0 = 200$, $H_1 = 1000$, D_0 is the controlled drop-rate setting, and D_1 is the restored published setting. Let

$$\theta_{hd} = \tau_{-I\to+I}(P_{hd}), \qquad h, d \in \{0, 1\}.$$

Table 29: **Targeted mechanistic follow-ups.** Cases were selected from the preceding protocol-sensitivity audit for diagnostic analysis. Selection of a case does not make it representative of the broader benchmark.

Case	Screening trigger on $-\Delta_{\mathrm{DP}}$	Planned diagnostic	Status
NIFTY/German	Controlled $-0.006 \rightarrow$ published -0.151	Separate the joint horizon and drop-rate change with an $H \times D$ design, then examine checkpoint-selection support within the long trajectory	Completed
FairGB/Bail	Controlled $-0.060 \rightarrow$ published-horizon comparison approximately -0.010	Cross-method trajectory follow-up for a case in which the published protocol change is restricted to training horizon	Completed
FairGB/Credit	Controlled $+0.014 \rightarrow$ published-horizon comparison $+0.081$	Planned companion to FairGB/Bail with an attribution shift in the opposite direction	Stopped at instrumentation-invariance gate

Table 30: **NIFTY/German factorial protocol decomposition on $-\Delta_{\mathrm{DP}}$.** Rows θ_{hd} report $\tau_{-I\rightarrow+I}(P_{hd})$ at horizon H_h and drop-rate setting D_d, where $H_0 = 200$, $H_1 = 1000$, and D_0 and D_1 denote the controlled and restored published drop-rate settings; the remaining rows report the contrasts defined above. Brackets give 95% paired hierarchical-bootstrap intervals, and bold values satisfy the resolved criterion. The short-horizon cells θ_{00} and θ_{01} are shared by both columns, which differ only in the long-horizon runs: the frozen protocol-comparison runs or the dedicated trajectory reruns (Appendix J.3). Within each column, $\theta_{11} - \theta_{00} = \Delta_H + \Delta_D$ up to rounding. *The interval excludes zero, but unit-level sign consistency is 0.733, below the 0.75 threshold; the corresponding rerun value is 0.767.

Quantity	Definition	Frozen long-horizon runs	Dedicated reruns
θ_{00}	(H_0, D_0)	-0.006 $[-0.042, +0.032]$	
θ_{01}	(H_0, D_1)	$+0.029$ $[-0.007, +0.073]$	
θ_{10}	(H_1, D_0)	-0.061 $[-0.128, +0.005]$	-0.043 $[-0.117, +0.032]$
θ_{11}	(H_1, D_1)	$\mathbf{-0.151}$ $[-0.225, -0.081]$	$\mathbf{-0.148}$ $[-0.221, -0.075]$
Δ_H	average horizon contrast	$\mathbf{-0.117}$ $[-0.184, -0.056]$	$\mathbf{-0.107}$ $[-0.176, -0.043]$
Δ_D	average drop-rate contrast	-0.027 $[-0.086, +0.017]$	-0.035 $[-0.098, +0.013]$
Γ_{HD}	interaction	-0.126 $[-0.220, -0.042]^*$	$\mathbf{-0.141}$ $[-0.240, -0.055]$
$\Delta_{H\mid D=1}$	$\theta_{11} - \theta_{01}$	$\mathbf{-0.180}$ $[-0.255, -0.107]$	$\mathbf{-0.177}$ $[-0.249, -0.104]$
$\theta_{11} - \theta_{00}$	endpoint shift	-0.145	-0.142

The average horizon-associated and drop-rate-associated contrasts are

$$\Delta_H = \frac{(\theta_{10} - \theta_{00}) + (\theta_{11} - \theta_{01})}{2}, \qquad \Delta_D = \frac{(\theta_{01} - \theta_{00}) + (\theta_{11} - \theta_{10})}{2},$$

and the interaction contrast is

$$\Gamma_{HD} = \theta_{11} - \theta_{10} - \theta_{01} + \theta_{00}.$$

The endpoint shift satisfies

$$\theta_{11} - \theta_{00} = \Delta_H + \Delta_D,$$

while Γ_{HD} describes whether the attribution change associated with one factor depends on the level of the other and is not a third additive component of the endpoint shift.

For the trajectory follow-up, the relevant contrast is the horizon-associated change under the restored published drop-rate setting,

$$\Delta_{H\mid D=1} = \theta_{11} - \theta_{01}.$$

Table 30 reports all four cells and the derived contrasts on $-\Delta_{\mathrm{DP}}$. Using the frozen long-horizon runs, the average horizon-associated contrast is $\Delta_H \approx -0.117$ and the horizon-associated contrast under the published drop-rate setting is $\Delta_{H\mid D=1} \approx -0.180$; both are resolved, and the dedicated reruns give -0.107 and -0.177, respectively. The drop-rate contrast is smaller and unresolved in both run sets ($\Delta_D \approx -0.027$ and -0.035). The interaction is negative in both run sets and its interval excludes zero in both, but its unit-level sign consistency (0.733 and 0.767) falls on either side of the 0.75 threshold, so its resolved status depends on the run set. We therefore do not interpret Γ_{HD} beyond noting that the horizon-associated contrast is larger in magnitude under D_1 than under D_0. Under validation-AUC selection, the corresponding published-setting estimate at H_1 is -0.020 and unresolved, so the horizon-associated shift in this case is selector-dependent.

Table 31: **Trajectory specifications for the completed mechanistic cases.** The capped selector operates on the stored long trajectory while restricting candidate checkpoints to those available by the short horizon. Fixed-checkpoint evaluations are descriptive trajectory diagnostics rather than additional primary experiments.

Method	Dataset	H_s	H_ℓ	Additional diagnostics
NIFTY	German	200	1000	BCE/AUC selectors; fixed checkpoints at 25, 50, 100, 150, 200, 300, 400, 600, 800, 1000
FairGB	Bail	200	1500	BCE/AUC selectors; fixed checkpoints at 25, 50, 100, 150, 200, 300, 500, 750, 1000, 1250, 1499

The subsequent analysis asks how $\Delta_{H|D=1}$ is related to expanding checkpoint-selection support within the stored long trajectory. The overall controlled-to-published NIFTY comparison is not interpreted as a horizon-only contrast because D also changes.

J.3 Trajectory Realization and Prefix Comparability

The selection-support decomposition is defined in Section 3.4. Here we document the case-specific trajectory construction used to instantiate it.

For both completed cases, extending the final training horizon does not alter the computations scheduled during the initial $H_s = 200$ epochs: neither implementation uses a learning-rate schedule whose earlier values depend on the requested final horizon, and the stochastic sampling logic does not condition on that horizon. We therefore use the first 200 epochs of each stored long trajectory as the candidate-checkpoint support for the capped selector σ_c.

This prefix comparability concerns execution logic rather than bitwise equality between independently executed trajectories. CUDA nondeterminism can produce residual numerical differences between a separately trained short trajectory and the corresponding prefix of a long trajectory. These differences are retained in R_{traj}; accordingly, R_{traj} is treated as a trajectory residual and not as an isolated training-horizon effect. By contrast, S_{sel} compares capped and full checkpoint selection on the same stored long trajectory and is therefore associated with expansion of checkpoint-selection support.

Table 31 summarizes the trajectory supports and fixed-checkpoint diagnostics.

J.4 Fixed-Checkpoint Trajectories

Figure 5 reports the matched intervention contrast at fixed checkpoints along the stored trajectories. For NIFTY/German, both the controlled drop-rate setting D_0 and restored published setting D_1 are shown. For FairGB/Bail, the trajectory extends through the published long horizon.

Under the restored published drop-rate setting, the NIFTY/German fairness contrasts move in the negative direction beyond the controlled checkpoint-support range. For FairGB/Bail, the fairness contrasts become most negative earlier in the trajectory and later move back toward zero. The two cases therefore do not support a common directional interpretation of longer training; the observed attribution change depends on the trajectory and on which checkpoints become available to the two intervention arms.

J.5 Three-Coordinate Decomposition

Table 32 reports R_{traj}, S_{sel}, and the corresponding arm-level selection-support terms for all three outcome coordinates. D_0 and D_1 denote the controlled and restored published NIFTY drop-rate settings, respectively. Bold entries satisfy the resolved reporting criterion.

For NIFTY/German under D_1, the dedicated trajectory analysis changes from approximately $+0.029$ at the short horizon to -0.148 under full long-trajectory selection on $-\Delta_{\mathrm{DP}}$; R_{traj} is near zero whereas $S_{\mathrm{sel}} \approx -0.177$. For FairGB/Bail, the corresponding contrast changes from approximately -0.060 to -0.019, with $R_{\mathrm{traj}} \approx -0.006$ and $S_{\mathrm{sel}} \approx +0.047$; the selection-support shift is associated primarily with the M^{+I} arm. These are case-specific diagnostic decompositions rather than estimates of a general training-horizon or checkpoint-selection effect. The accompanying result

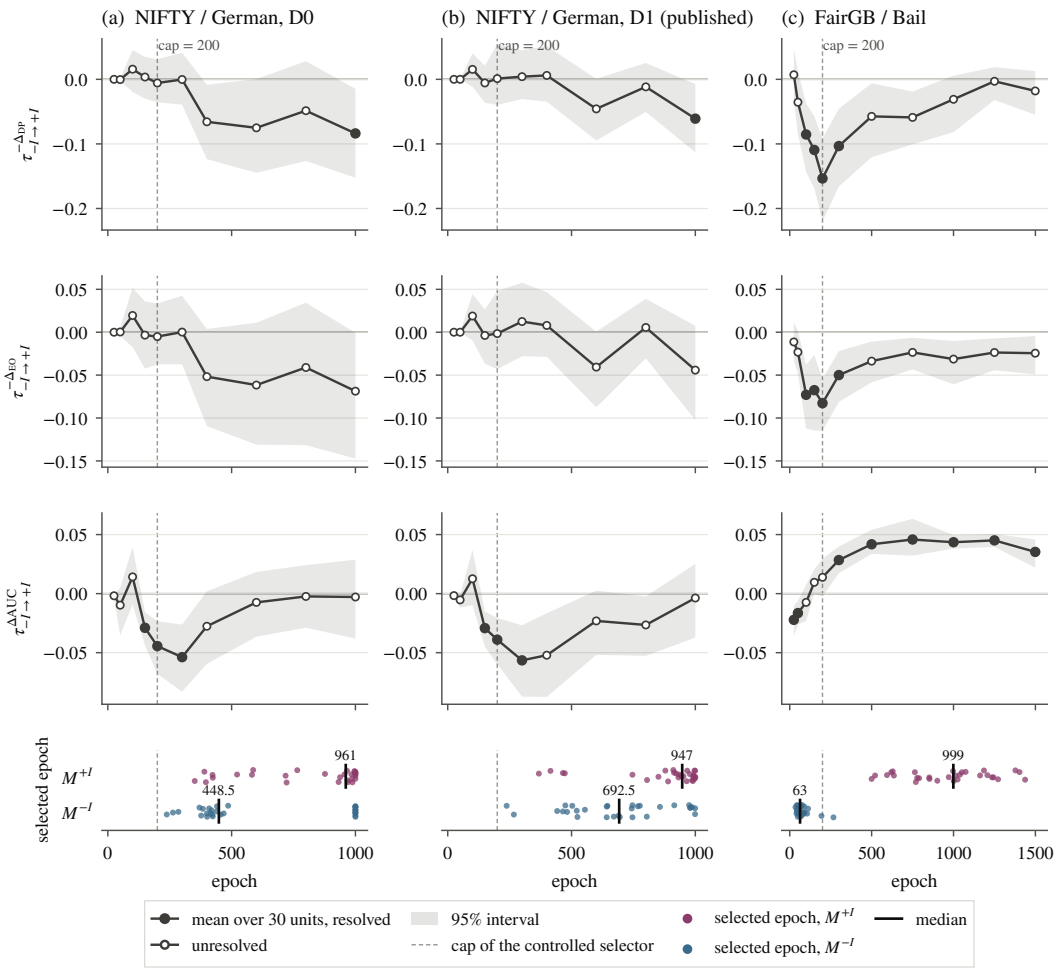


Figure 5: **Fixed-checkpoint trajectories in the mechanistic case studies.** Rows report $\tau_{-I\to+I}(P)$ for $-\Delta_{\mathrm{DP}}$, $-\Delta_{\mathrm{EO}}$, and $\Delta\mathrm{AUC}$, followed by the selected checkpoint epochs for M^{+I} and M^{-I}. The vertical line marks the $H = 200$ checkpoint-support cap; shading denotes 95% bootstrap intervals.

Table 32: **Selection-support decomposition across outcome coordinates.** The bridge satisfies $\tau_{\mathrm{cap}} = \tau_{H=200} + R_{\mathrm{traj}}$ and $\tau_{\mathrm{full}} = \tau_{\mathrm{cap}} + S_{\mathrm{sel}}$. Bold values satisfy the resolved criterion.

Case	Coordinate	$\tau_{H=200}$	R_{traj}	τ_{cap}	S_{sel}	τ_{full}
NIFTY/German D0	$\Delta\mathrm{AUC}$	-0.031	+0.003	-0.028	+0.043	+0.015
	$-\Delta_{\mathrm{DP}}$	-0.006	-0.005	-0.011	-0.032	-0.043
	$-\Delta_{\mathrm{EO}}$	+0.004	-0.000	+0.004	-0.030	-0.026
NIFTY/German D1	$\Delta\mathrm{AUC}$	-0.023	-0.000	-0.023	**+0.072**	**+0.049**
	$-\Delta_{\mathrm{DP}}$	+0.029	-0.000	+0.029	**-0.177**	**-0.148**
	$-\Delta_{\mathrm{EO}}$	+0.029	+0.000	+0.030	**-0.152**	**-0.122**
FairGB/Bail	$\Delta\mathrm{AUC}$	**+0.019**	+0.001	**+0.020**	**+0.044**	**+0.064**
	$-\Delta_{\mathrm{DP}}$	-0.060	-0.006	-0.066	**+0.047**	-0.019
	$-\Delta_{\mathrm{EO}}$	**-0.065**	-0.006	**-0.072**	**+0.058**	-0.013

Notes. D0 and D1 denote the controlled and restored published NIFTY drop-rate settings. R_{traj} retains differences between the separately executed short run and the capped prefix of the long run and is not interpreted as an isolated horizon effect. The mechanistic cases are targeted diagnostics and are not included in benchmark aggregate counts.

file contains derived decomposition-term–coordinate records from these targeted cases; these rows are not independent experiments.

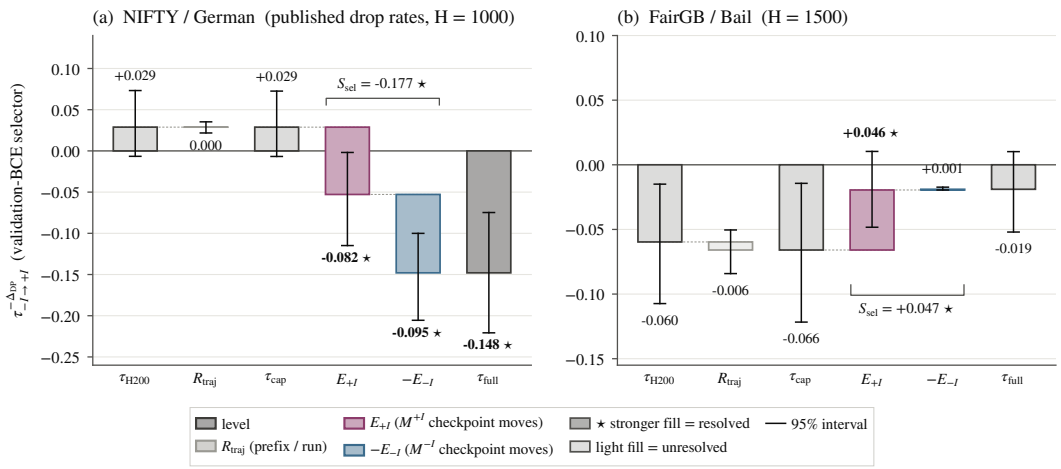


Figure 6: **Selection-support decomposition on $-\Delta_{\mathrm{DP}}$.** Bars show $\tau_{H=200}$, R_{traj}, τ_{cap}, the arm-level terms E_{+I} and $-E_{-I}$, and τ_{full}; $S_{\mathrm{sel}} = E_{+I} - E_{-I}$. Error bars denote 95% bootstrap intervals. Values for all three coordinates are in Table 32.

Table 33: **FMP component-level analysis grid.** This targeted component study is reported separately from the primary binary-intervention evaluation.

Component	Evaluated values
Datasets	Pokec-z, Pokec-n
Fairness strength λ_1	$5, 30$
Propagation-related strength λ_2	$0.01, 20$
Checkpoint selector	Last, validation BCE, validation AUC
Outcome coordinates	$\Delta\mathrm{AUC}, -\Delta_{\mathrm{DP}}, -\Delta_{\mathrm{EO}}$

Forcing both arms to a common checkpoint, defined as the epoch minimizing the average validation BCE of the two arms, reduces the NIFTY/German contrasts under D_1 to -0.109 on $-\Delta_{\mathrm{DP}}$ and -0.091 on $-\Delta_{\mathrm{EO}}$, both still resolved, and to $+0.023$ on $\Delta\mathrm{AUC}$, which is no longer resolved; under D_0 all three contrasts remain unresolved, and no sign changes. For FairGB/Bail, test scores were stored only at eleven grid epochs, so the same check is approximate and is not interpreted.

K FMP Component-Level Attribution

FMP is analyzed separately from the primary binary-intervention benchmark because its released implementation exposes an interpretable sequence of component boundaries,

$$B \to M^{\mathrm{base}} \to M^{\mathrm{prop}}_{\lambda_2} \to M^{\mathrm{prop+fair}}_{\lambda_1,\lambda_2}.$$

This targeted analysis asks whether a complete-method change can be localized to the explicit fairness-correction stage or instead arises earlier in the component path. FMP is eligible for this analysis because the released implementation provides a propagation-only state that can be compared with the full model while holding the propagation configuration fixed. The eligibility rule and parameter-grid selection rule were frozen before evaluating the component contrasts. Among the supported datasets, we retain Pokec-z and Pokec-n; NBA is excluded because the released evaluation uses test information for validation and therefore does not support the leakage-free checkpoint-selection analysis used here.

We define three successive component contrasts:

$$\tau_{B\to\mathrm{base}} = \mathbb{E}_\omega\left[Y_\omega(M^{\mathrm{base}}) - Y_\omega(B)\right],$$

$$\tau_{\mathrm{base}\to\mathrm{prop}}(\lambda_2) = \mathbb{E}_\omega\left[Y_\omega(M^{\mathrm{prop}}_{\lambda_2}) - Y_\omega(M^{\mathrm{base}})\right],$$

and

$$\tau_{\mathrm{prop}\to\mathrm{fair}}(\lambda_1,\lambda_2) = \mathbb{E}_\omega\left[Y_\omega(M^{\mathrm{prop+fair}}_{\lambda_1,\lambda_2}) - Y_\omega(M^{\mathrm{prop}}_{\lambda_2})\right].$$

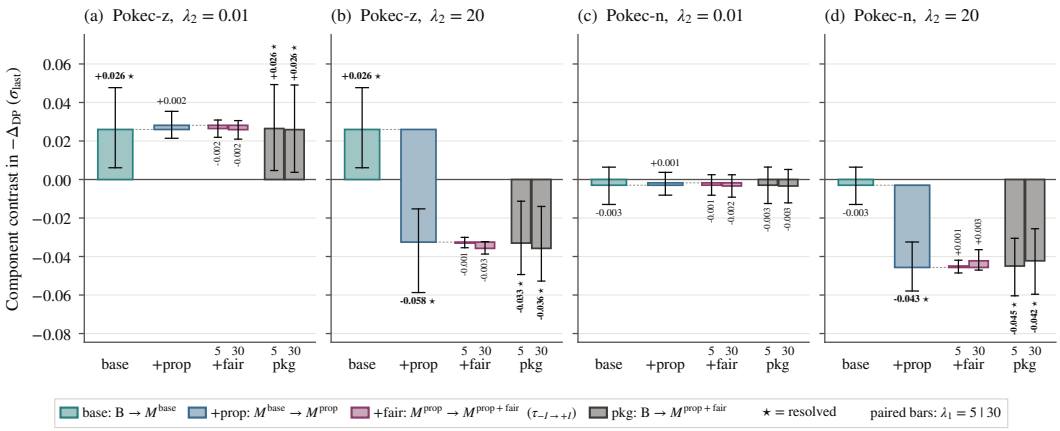


Figure 7: **Component-level attribution in FMP.** Bars report the baseline-to-base, propagation, explicit fairness-correction, and complete-method contrasts for $-\Delta_{\mathrm{DP}}$ under the last-checkpoint selector. Error bars denote 95% bootstrap intervals.

The corresponding path totals satisfy

$$\tau_{\mathrm{base}\to\mathrm{prop+fair}} = \tau_{\mathrm{base}\to\mathrm{prop}} + \tau_{\mathrm{prop}\to\mathrm{fair}}, \qquad \tau_{B\to\mathrm{prop+fair}} = \tau_{B\to\mathrm{base}} + \tau_{\mathrm{base}\to\mathrm{prop+fair}}.$$

Here M^{base} is the released FMP model with $\lambda_1 = \lambda_2 = 0$, which disables both propagation and the fairness correction and reduces the model to an MLP. Under this protocol, the complete FMP has larger demographic-parity disparity than B in six of the eight configurations under the last-checkpoint selector, including all four with $\lambda_2 = 20$, where the difference is resolved. The finding therefore concerns the incremental contribution of the explicit correction in this regime, not FMP's published fairness gains.

The external baseline B in this analysis is a vanilla GCN trained under the FMP protocol (300 epochs, last-checkpoint selection) and therefore differs from the reference used in the primary evaluation; $\tau_{B\to\mathrm{base}}$ should not be compared directly with $\tau_{B\to-I}$ in Table 1. These are algebraic component contrasts. In particular, $\tau_{\mathrm{prop}\to\mathrm{fair}}$ is interpreted as the incremental contribution associated with adding the explicit fairness correction conditional on the same propagation configuration; it does not by itself identify a mechanism.

We evaluate $\lambda_1 \in \{5, 30\}$ and $\lambda_2 \in \{0.01, 20\}$, the smallest and largest positive values in the released FMP hyperparameter sweep, chosen by a rule fixed before any component contrast was computed; because the released code specifies no dataset-specific setting, we report the component path at these endpoints rather than at a tuned value, under the last checkpoint, validation-BCE selection, and validation-AUC selection. The propagation contrast is evaluated with the fairness branch disabled, whereas the fairness-correction contrast compares states with the same λ_2 and differs by addition of the explicit fairness component. Because λ_2 also enters internal scaling used by the fairness correction, it is not interpreted as a perfectly isolated propagation-only factorial factor; the clean comparisons are the propagation increment with the fairness branch disabled and the subsequent fairness increment conditional on that propagation configuration.

K.1 Primary Component Decomposition

Figure 7 reports the component path on $-\Delta_{\mathrm{DP}}$ under the last-checkpoint selector. Across the evaluated configurations, substantial complete-method changes can arise at the base or propagation stages while the subsequent explicit fairness-correction increment remains small. For example, on Pokec-z with $\lambda_2 = 0.01$, the baseline-to-base contrast is approximately $+0.026$ and resolved, whereas the subsequent fairness-correction increments are approximately -0.002.

Under the last-checkpoint selector, none of the 24 explicit fairness-correction contrasts across the eight dataset–configuration combinations and three outcome coordinates is resolved, and every absolute point estimate is below 0.005. This pattern does not imply that the complete FMP has no

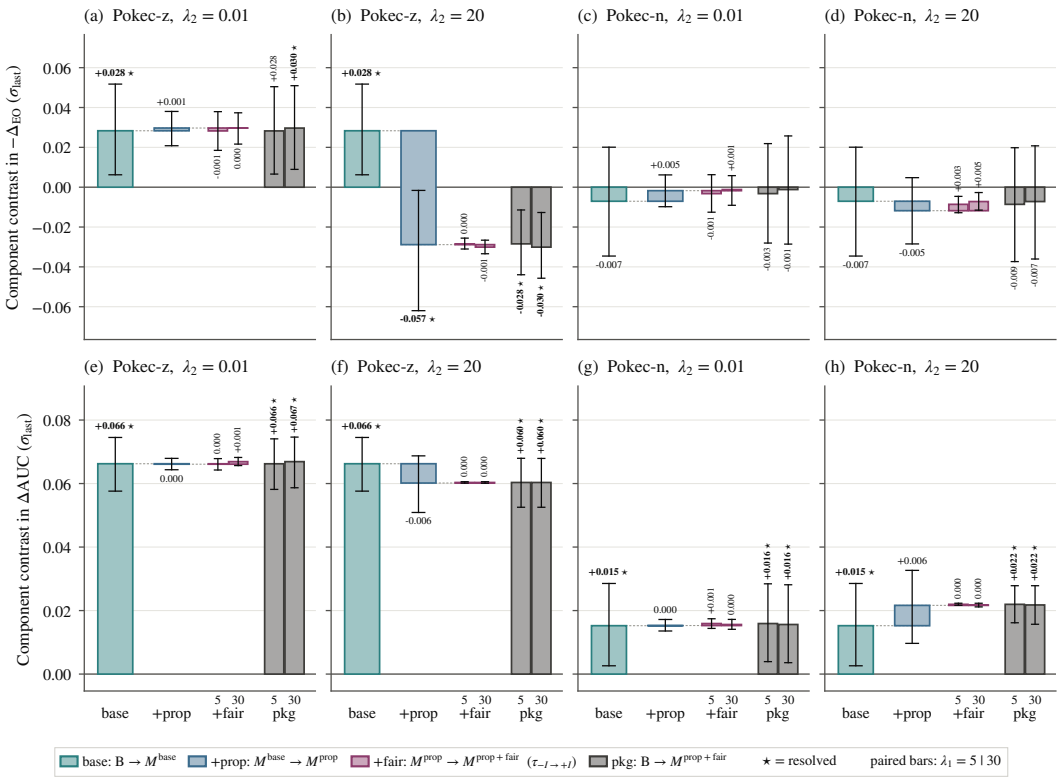


Figure 8: **FMP component-level attribution for $-\Delta_{\mathrm{EO}}$ and $\Delta\mathrm{AUC}$.** Bars show the same component boundaries as Figure 7 under the last-checkpoint selector; error bars denote 95% bootstrap intervals.

fairness-related changes; rather, within this component construction, the larger observed differences can already be present before the explicit fairness-correction stage.

K.2 Additional Outcome Coordinates

Figure 8 extends the same component decomposition to $-\Delta_{\mathrm{EO}}$ and $\Delta\mathrm{AUC}$. The equal-opportunity results likewise show settings in which a larger component change occurs before the explicit fairness correction. For Pokec-z with $\lambda_2 = 20$, for example, the propagation increment on $-\Delta_{\mathrm{EO}}$ is approximately -0.057, whereas the subsequent fairness-correction increments remain close to zero. On $\Delta\mathrm{AUC}$, all eight complete-method contrasts under the last-checkpoint selector are positive and resolved, ranging from approximately $+0.016$ to $+0.067$, while the corresponding explicit fairness-correction increments are approximately $+0.0001$ to $+0.0008$.

K.3 Selector Robustness

We finally evaluate whether the small explicit fairness-correction increment is specific to the last-checkpoint rule. Figure 9 reports $\tau_{\mathrm{prop}\to\mathrm{fair}}(\lambda_1, \lambda_2)$ under the last, validation-BCE, and validation-AUC selectors over the full evaluated grid.

Across the 72 dataset–configuration–selector–coordinate evaluations, no explicit fairness-correction contrast is resolved. The largest absolute point estimate is approximately 0.005. Some percentile intervals exclude zero, but these contrasts remain unresolved because the reporting rule additionally requires $|\widehat{\tau}| \geq 0.01$. Thus, within the evaluated FMP grid, the small incremental contribution associated with the explicit fairness correction is not specific to one checkpoint-selection rule.

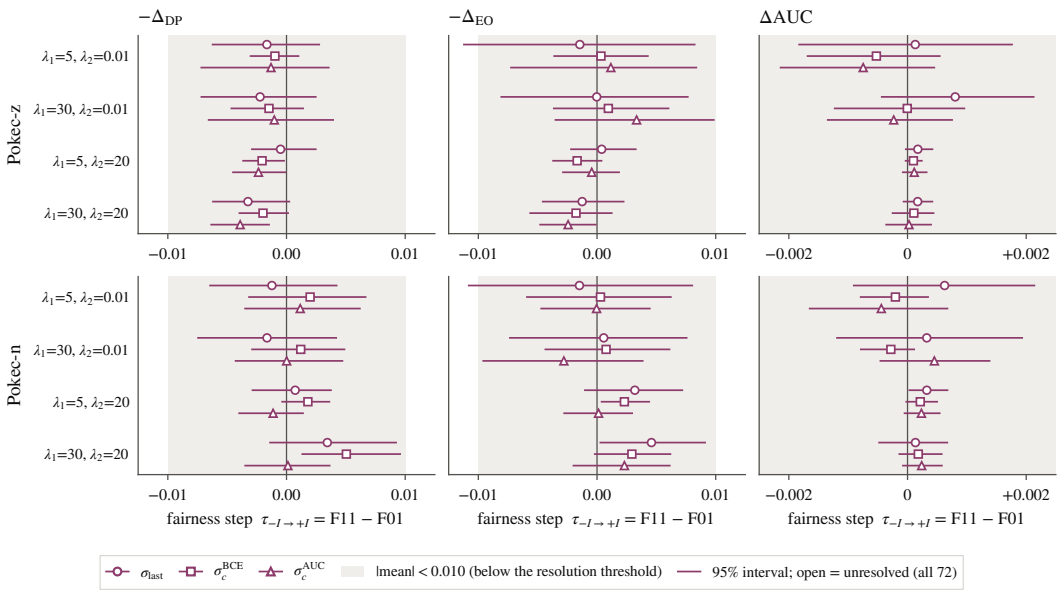


Figure 9: **FMP explicit fairness-correction contrast across checkpoint selectors.** Points report $\tau_{\text{prop}\to\text{fair}}(\lambda_1, \lambda_2)$ under the three checkpoint rules; error bars denote 95% bootstrap intervals. None of the 72 selector–configuration–coordinate evaluations satisfies the resolved reporting criterion.

The accompanying artifact provides the complete derived component table in `results/5_component_case_study_FMP.csv`. Its 270 rows are derived component–parameter–selector–coordinate records and are not interpreted as 270 independent experiments.