SUCCESSFUL CONSTRAINT OPTIMIZATION NEED NOT CERTIFY STRUCTURAL FIDELITY: A VERIFICATION GAP IN GRAPH STRUCTURE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph structure learning methods often adapt graph structure from data while incorporating predefined graphs, structural priors, or constraint-based objectives. When a learned graph is claimed to realize a specific external structural target, however, optimization success raises a separate verification question: do the optimized and reported quantities actually certify fidelity to that target? We study this question in a controlled declared-target setting. An external graph A^* guides a warmup phase; thereafter, the main non-oracle training arms continue without allowing A^* to influence graph updates, while A^* is retained for structural audit. Within the Hinge runs on the primary traffic networks, the residual proxy decreases by more than 98% from the warmup boundary and approaches zero. At the final epoch, paired Hinge–ReLU task-error differences remain small, whereas evaluation-only structural audits show lower target fidelity under Hinge than under ReLU. This lower-target-fidelity pattern replicates across four amortized graph learners and multiple traffic networks. We further characterize when surrogate near-optimality can certify proximity to the declared target and isolate two distinct failure points: the constraint interface may lack sufficient target-discriminating information, or the realized graph updates may move in a structurally unfavorable direction. An A^*-dependent oracle gradient intervention consistently improves local structural quantities but does not reliably restore global edge ranking. This recovery boundary motivates limiting post-warmup deviation from a high-fidelity warmup anchor without requiring A^* for subsequent non-oracle updates. The resulting guardrail substantially limits structural drift relative to unconstrained continued training; under a controlled fixed-support shift, it retains more structural fidelity than unconstrained adaptation while adapting better than freezing. Thus, task-side behavior and successful surrogate optimization do not, by themselves, certify fidelity to a declared external target. Such claims require direct structural evidence or a demonstrated certificate linking the reported quantities to that target.

1 INTRODUCTION

Graph structure learning (GSL) often combines adaptive topology with predefined graphs, domain priors, or constraint-based objectives. Such structure can play different roles, from predictive inductive bias to an explicitly declared structural target. Only the latter makes fidelity to the external structure an empirical claim and therefore creates a stronger evidential obligation: favorable task performance and successful constraint-guided optimization do not, by themselves, establish that the learned graph remains faithful to the specified target. We study when successful optimization through a constraint interface can provide a structural certificate and where this implication can fail.

We isolate this question through a declared-target training contract. An external graph A^* anchors the learner during warmup; after warmup, the main non-oracle training arms continue through the task and constraint interfaces without allowing A^* to influence graph updates. The target is retained for structural audit and explicitly labeled diagnostic or oracle controls. This setting isolates the case in which fidelity to an external structure is itself part of the claim, rather than assuming that every prior-guided graph learner makes such a claim. This contract does not assume that the post-

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.

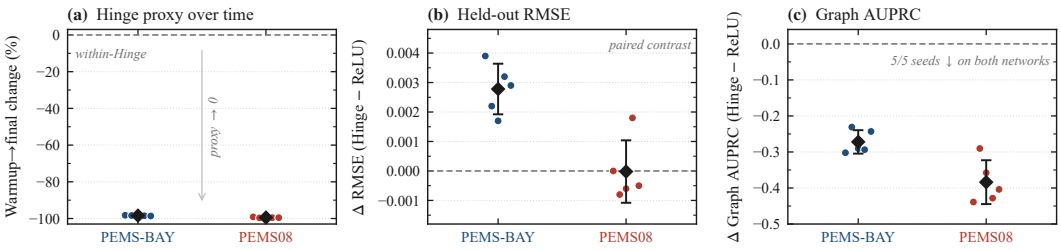


Figure 1: **Proxy optimization, predictive error, and structural fidelity diverge.** Results are paired across five seeds on PEMS-BAY and PEMS08. The left panel shows the within-Hinge warmup-to-final proxy change; the other panels show final-epoch Hinge-minus-ReLU contrasts in RMSE and Graph AUPRC. Diamonds and whiskers denote mean ± sample SD. The proxy falls sharply, RMSE differences remain small, and Graph AUPRC is lower under Hinge in every paired seed.

warmup constraint interface identifies the target; whether success at that interface provides evidence of continued fidelity to A^* is precisely the verification question.

Prediction–structure mismatch is already documented: prediction-useful graphs can differ from known relations, and strong predictive models can recover underlying interactions poorly (Zügner et al., 2021; Ozen et al., 2025; Dubbeldam et al., 2026; Shi et al., 2026). Relatedly, Manenti et al. show that point-prediction losses need not recover the distribution of a latent graph (Manenti et al., 2025). Existing mismatch results therefore establish that predictive success and structural recovery can diverge. They do not by themselves answer the certificate question studied here: under what conditions does success at a constraint-guided interface provide valid evidence of fidelity to a particular declared target? The resulting verification gap is directly observable. On PEMS-BAY and PEMS08, the residual proxy within the Hinge runs decreases by more than 98% from the warmup boundary and approaches zero. At the final epoch, the Hinge–ReLU task-error differences remain small, while Graph AUPRC is lower under Hinge in every paired seed on both networks (Figure 1). This three-way dissociation separates task behavior, surrogate success, and target fidelity: neither of the first two, by itself, certifies target fidelity.

The gap exposes two auditable failure points (Figure 2). **Calibration** asks whether the constraint interface contains enough target-discriminating information to separate A^* from structurally different alternatives; **Route** asks whether the realized graph updates move in a structurally favorable direction. At a fixed learned state, we characterize when surrogate near-optimality can certify structural proximity and use controlled interventions to probe both failure points. An A^*-dependent route correction consistently improves local structural quantities but yields mixed global-ranking outcomes, revealing a local-to-global recovery boundary. This boundary motivates protecting a previously audited high-fidelity structure during adaptation: a warmup-anchor guardrail limits post-warmup structural drift in a real traffic setting and, under a controlled fixed-support shift, retains more structural fidelity than unconstrained adaptation while adapting better than freezing.

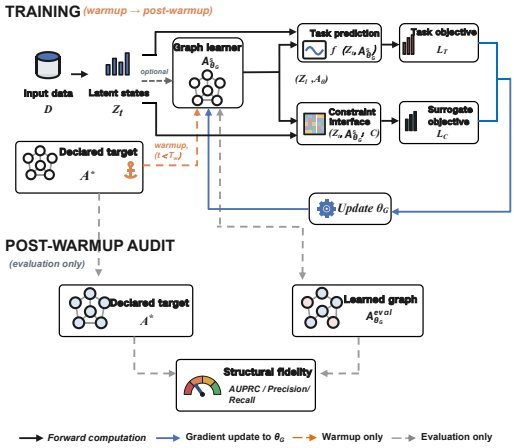


Figure 2: **Training and evaluation-only audit interface.** Black arrows denote forward computation, blue arrows denote graph-parameter updates, orange dashed arrows denote warmup-only access to A^*, and gray dashed arrows denote evaluation-only audit paths. After warmup, ordinary training updates no longer access A^*.

Our contributions follow this diagnostic sequence:

D1. Discovery. We establish a three-way dissociation among surrogate success, task-side behav-

ior, and target fidelity on the primary traffic benchmarks. We further replicate the finding of lower final target fidelity under Hinge than under ReLU across four amortized graph learners and multiple traffic networks.

D2. Diagnosis. At a fixed learned state, we characterize when surrogate near-optimality can certify structural proximity and diagnose two distinct failure points, Calibration and Route, through controlled interventions and directional gradient audits.

D3. Boundary and implication. Correcting an adverse local route does not reliably restore global edge ranking; a warmup-anchor guardrail limits post-warmup structural drift while preserving useful adaptation in the tested shift without requiring post-warmup access to A^*.

2 Related Work

Graph structure learning and structural claims. Graph structure learning jointly adapts graph topology and downstream representations (Franceschi et al., 2019; Chen et al., 2020; Jin et al., 2020; Shang et al., 2021). Learned relations may also coexist with fixed graph supports or optional structural priors (Wu et al., 2019; Shang et al., 2021). Related traffic and time-series methods update or combine graph estimates through iterative graph refinement, hybrid graph fusion, prior-guided graph losses, or non-end-to-end adaptive graph learning, respectively (Zhang et al., 2022a; Guo et al., 2023; Wang et al., 2024; Xu et al., 2025). Learned graphs can also carry different semantic interpretations: sensor dependencies in anomaly detection (Deng & Hooi, 2021; Zhang et al., 2022b), process relations used for root-cause interpretation (Chen et al., 2025), latent relational dependencies among concepts in knowledge tracing (Jia et al., 2026), or directed predictive relevance in forecasting (Abdulsahib et al., 2026). Our verification question arises specifically when fidelity to a specified external structure is itself part of the claim; it does not impose that obligation on learned graphs used only as task-adaptive representations.

Auditing and recovering learned structure. Beyond downstream performance, recent work has evaluated learned graphs as structural objects in their own right. In multivariate time-series models, Zügner et al. (2021) assess graph inference alongside forecasting through interventions that disable graph learning or supply ground-truth relations, while Ozen et al. (2025) directly audit GDN's graph-learning module and find that its end-to-end performance arises largely from the downstream model rather than the learned graph itself. With known reference structure, Dubbeldam et al. (2026) evaluate whether graph-based tabular models recover ground-truth feature interactions, showing that predictive modeling can succeed without recovering the intended interaction graph. In industrial forecasting, Shi et al. (2026) further identify a setting in which a longer look-back window improves prediction while degrading learned-graph accuracy. Relatedly, Wienöbst et al. (2026) show in score-based causal discovery that, when the scoring criterion fails to identify the true graph, successful Gaussian-BIC optimization can yield graphs that score better than the ground truth while structural recovery still suffers. Theoretically, Manenti et al. (2025) show that minimizing a point-prediction loss does not in general guarantee recovery of the latent-graph distribution, and derive losses under which predictive optimality and latent-structure learning can be achieved jointly. Together, these studies establish that predictive success and structural recovery are distinct objectives and need not coincide. Our setting asks a different evidential question: once a specific external structure is declared as the fidelity target and the main non-oracle post-warmup updates continue through task and constraint interfaces without allowing that target to influence them, does successful constraint-guided optimization provide valid evidence that the declared structure remains realized?

Verification and surrogate validity. More generally, success on an optimized or reported quantity supports a stronger target claim only when that quantity is appropriately linked to the target property. Classical surrogate-calibration theory formalizes this issue by asking when optimizing a tractable surrogate preserves performance under the target loss (Bartlett et al., 2006). Related concerns arise in explainability, where commonly used fidelity measures can misalign with the intended notion of faithfulness (Zheng et al., 2024; 2025), and where successful interventions or existing validity criteria need not by themselves establish the intended mechanism or explanation (Makelov et al., 2024; Méloux et al., 2025). We bring this evidence-validity question to learned graph structure: when fidelity to a specified external graph is itself part of the claim, we formulate a structural-certificate

question—when is surrogate near-optimality sufficient to certify proximity to the declared target? Such validity can fail when the observable constraint interface cannot distinguish the declared target from structurally different candidates, a condition formalized in Section 5.1.

Constraint interfaces and update routes. Constraint-based learning provides a general mechanism for translating symbolic or structural specifications into trainable objectives. Differentiable-logic frameworks instantiate this idea for logical knowledge (Xu et al., 2018; Fischer et al., 2019; Badreddine et al., 2022), while structure-learning methods encode graph properties such as acyclicity through continuous constraints (Zheng et al., 2018; Yu et al., 2019). Subsequent work exposes optimization pathologies (Wei et al., 2020), incorporates ancestral constraints (Ban et al., 2025), or derives analytic DAG constraints designed to mitigate gradient vanishing (Zhang et al., 2025). These works primarily address how structural specifications can be represented and optimized. Our question instead concerns whether success at the resulting interface provides valid evidence of fidelity to the declared graph. Multi-objective methods study interactions among training objectives, for example through Pareto-compatible updates or modifications of conflicting gradients (Desideri, 2012; Sener & Koltun, 2018; Yu et al., 2020). Our Route analysis instead asks whether the realized graph update is structurally favorable relative to A^*, thereby separating objective–objective interaction from update–structure alignment. Together, these distinctions separate the mechanics of constraint optimization from the evidential question of structural realization.

3 Problem Setup

3.1 Constraint-Guided Structure Learning

Let $\mathcal{V}$ be a node set of size N, $\mathcal{D}$ an observed sequence dataset, and $\ell_{\theta_G} \in \mathbb{R}^{N\times N}$ the deterministic base edge scores produced by the graph learner. Let $A^s_{\theta_G} \in [0,1]^{N\times N}$ denote the training-time adjacency constructed from these scores and used by the forward and constraint interfaces, and let $A^{\mathrm{eval}}_{\theta_G} \in [0,1]^{N\times N}$ denote the deterministic adjacency induced by the current graph parameters for structural evaluation. We further denote the learned node states by Z_{θ_E}, the task-head parameters by θ_H, and the constraint specification by $\mathcal{C}$.

For stochastic graph parameterizations, $A^s_{\theta_G}$ may include training-time sampling noise; for offset-based forms, it may also reflect a transient constraint-induced modification of the edge scores. In contrast, $A^{\mathrm{eval}}_{\theta_G}$ is computed without sampling noise or transient form-specific constraint offsets and represents the adjacency encoded by the current learned graph parameters.

Using the objects defined above, the loss-form interface at epoch t is

$$\mathcal{J}_t = \mathcal{L}_{\mathrm{task}} + \mathbf{1}[t \geq T_w]\lambda\mathcal{L}_c + \mathbf{1}[t < T_w]\lambda_{w,t}\mathcal{L}_{\mathrm{warm}}. \tag{1}$$

Here, $\lambda \geq 0$ is the post-warmup constraint-loss weight and $\lambda_{w,t} \geq 0$ the warmup weight, with $\lambda_{w,t} = \lambda_w(1 - t/T_w)$ when warmup is used and $\lambda_{w,t} = 0$ otherwise.

Equation (1) applies to Hinge and other loss-based interfaces. In offset-based forms such as ReLU-zscore, the constraint signal modifies the edge scores before $A^s_{\theta_G}$ is constructed rather than entering through an additive $\lambda\mathcal{L}_c$ term; Appendix B gives the form-specific details.

For all main non-oracle arms, A^* influences training only through the warmup objective for $t < T_w$. After warmup, optimization proceeds through the task and applicable constraint interfaces without allowing A^* to influence ordinary updates. The target remains available for structural evaluation and diagnostics, including SDA, provided that this information is not fed back into training. Post-warmup target-dependent interventions— including the gradient-projection oracle, Support Oracle, and positive structural-target controls—are explicitly labeled oracle or direct-target controls and are excluded from this non-oracle contract.

All structural audits use $A^{\mathrm{eval}}_{\theta_G}$, excluding training-time sampling noise and transient constraint-induced modifications. Graph AUPRC is computed from its off-diagonal soft scores, whereas Precision, Recall, and Density use the corresponding parameterization-specific thresholded graph. Further implementation details are given in Appendices A–B.

3.2 Semantics, Metrics, and Scope

Traffic and EdNet instantiate different structural semantics. In traffic, reference edges encode spatial connectivity or proximity and do not impose an ordering on latent states. The Hinge objective is therefore a residual-based training surrogate, not a direct road-identification or physical-law objective. Accordingly, low Hinge loss is not, by construction, a measure of adjacency fidelity; whether its optimization provides evidence of fidelity to a declared target therefore requires a separate verification argument. EdNet provides a complementary setting with more direct alignment between constraint and structural semantics: a directed relation $i \to j$ is interpreted through an ordering such as $z_i \geq z_j$.

We use Graph AUPRC on $A^{\mathrm{eval}}_{\theta_G}$ as the primary soft-ranking measure of fidelity to A^*, with Precision, Recall, and Density computed on its thresholded counterpart. LVR measures state-level conformity and the traffic residual proxy measures surrogate optimization; neither is treated as an adjacency-fidelity audit. All claims of fidelity are relative to the declared reference A^* and do not imply that A^* is uniquely task-optimal or semantically valid.

Unless otherwise noted, all comparisons use five paired random seeds. We report within-seed contrasts, mean $\pm$ sample standard deviation ($\mathrm{ddof} = 1$), and direction counts; we make no large-sample significance claims. Full protocols and the precise scope of each intervention appear in Appendices A and F.

4 Constraint–Structure Dissociation [D1]

We first test whether successful surrogate optimization, together with small observed task-side differences, provides sufficient evidence that the learned graph remains faithful to the declared target.

4.1 Primary Three-Way Dissociation

On PEMS-BAY and PEMS08, the learned-edge-weighted Hinge proxy decreases within the Hinge runs from $0.306/0.429$ at the warmup boundary to $0.005/0.002$ at the final epoch, a reduction of more than 98% on both networks. At the final epoch, however, using ReLU as a matched alternative constraint form rather than a structural gold standard, Hinge yields lower Graph AUPRC in every paired seed, with Hinge-minus-ReLU differences of -0.272 ± 0.033 and -0.384 ± 0.061, while the corresponding held-out RMSE differences remain small at approximately $+0.003$ and 0.000 (Figure 1). These quantities form an evidential dissociation: the within-Hinge change establishes surrogate optimization, whereas the matched final-epoch contrasts assess task-side behavior and target fidelity. The RMSE comparison is descriptive; no equivalence margin was prespecified.

EdNet provides a complementary order-domain state-conformity–structure stress test rather than a task-aware replication of the three-way result. Relative to ReLU-zscore, Hinge reduces the reference-edge violation rate from 0.498 to 0.047, while Graph AUPRC decreases from 0.880 to 0.427, yielding an analogous state-conformity–structure reversal under prerequisite-order semantics. Because task performance is weak and learned-graph dependence is not established in this setting, we use EdNet only as structural stress-test evidence.

4.2 Replication Across Parameterizations and Networks

The lower target-fidelity pattern is not confined to a single amortized graph parameterization or traffic network. On METR-LA, the Hinge-minus-ReLU Graph-AUPRC difference is negative in all five paired seeds for four amortized graph learners—Inner-product, MTGNN, Graph WaveNet, and EdgeMLP. The same direction is reproduced with EdgeMLP on PEMS-BAY and PEMS08. Thus, lower target fidelity under Hinge persists across substantially different adaptive-adjacency parameterizations and multiple traffic networks. FreeLogit does not show this pattern and therefore marks an observed parameterization boundary rather than an additional replication.

When the final edge budget is matched by magnitude pruning, structural fidelity remains substantially higher than under Hinge, showing that the observed degradation is not explained by sparsity alone. In a positive structural-target control, the pipeline recovers the explicitly supplied target, ruling out generic pipeline incapacity as an explanation. The dissociation also persists through the

preregistered 500-epoch horizon, ruling out a short-lived post-warmup transient within the tested horizon. Taken together, these results show that the lower target-fidelity pattern associated with the constraint–structure dissociation is reproducible across the tested learners and networks and is not explained by a single parameterization, edge-budget effect, or short-lived training transient. Complete paired contrasts and control details are reported in Appendix Table 3 and Appendix F.

5 CALIBRATION × ROUTE [D2]

D1 establishes that surrogate optimization, task behavior, and structural fidelity can diverge. We now ask where the evidential link from optimization success to structural fidelity can break. We separate two auditable failure points. Calibration asks whether, at a fixed learned state, the constraint interface contains enough target-specific information to distinguish the declared structure from structural alternatives. Route asks whether the realized graph update moves in a structurally favorable direction. Figure 2 makes this separation explicit by distinguishing training-time update paths from the evaluation-only structural audit.

5.1 CONSTRAINT-TO-STRUCTURE CALIBRATION

For surrogate near-optimality to serve as a structural certificate, it must exclude structures that remain separated from the declared target. Let $\mathcal{A}$ be the feasible adjacency set, and let $D(A, A^*) \geq 0$ be a claim-relative structural risk whose zero set consists exactly of the structures regarded as faithful to A^*. This set is $\{A^*\}$ under exact fidelity, but may more generally be an equivalence class $\mathcal{E}(A^*)$. For fixed Z, define $\mathcal{L}_c^*(Z) := \inf_{A\in\mathcal{A}} \mathcal{L}_c(A; Z)$, and assume that this infimum is finite.

Definition 1 (Constraint-to-Structure Calibration). The surrogate is structure-calibrated on $(Z, \mathcal{A}, A^*, D)$ if, for every $\epsilon > 0$ for which the set in the infimum below is nonempty,

$$\Psi_Z(\epsilon) := \inf_{\substack{A\in\mathcal{A} \\ D(A,A^*)\geq\epsilon}} [\mathcal{L}_c(A; Z) - \mathcal{L}_c^*(Z)] > 0.$$

The margin $\Psi_Z(\epsilon)$ separates surrogate optimality from every structure that remains at least ϵ away from the declared target. When this margin collapses, arbitrarily small surrogate excess remains compatible with structurally separated graphs.

Proposition 1 (Structural Certificate Characterization). For fixed Z, Definition 1 holds if and only if every sequence $\{A_n\} \subseteq \mathcal{A}$ satisfies

$$\mathcal{L}_c(A_n; Z) - \mathcal{L}_c^*(Z) \to 0 \quad \Longrightarrow \quad D(A_n, A^*) \to 0.$$

At any admissible tolerance ϵ, the same margin yields the finite certificate $\mathcal{L}_c(A; Z) - \mathcal{L}_c^*(Z) < \Psi_Z(\epsilon) \Rightarrow D(A, A^*) < \epsilon$. Appendix E.1 gives the proof.

The characterization exposes an information-level obstruction: optimization cannot recover structural distinctions that are absent from the constraint interface.

Lemma 1 (Interface-Equivalence Limit). Suppose that $\mathcal{L}_c(A; Z) = h_Z(T_Z(A))$, $A^* = A_0 \in \arg\min_{A\in\mathcal{A}} \mathcal{L}_c(A; Z)$, $D(A_1, A_0) > 0$, and $T_Z(A_1) = T_Z(A_0)$. Then

$$\Psi_Z(D(A_1, A_0)) = 0.$$

Thus, if a structurally separated alternative is indistinguishable from the declared target at the constraint interface, surrogate optimality alone cannot provide a structural certificate. The proof and a corresponding indistinguishability result are given in Appendix E.2.

The local geometry of the implemented learned-edge-weighted Hinge objective is consistent with this information-side limitation. With states and total edge mass fixed, the objective favors residual-compatible edges rather than target edges per se (Appendix E.3). Its local preference is therefore determined by residual cost, which need not preserve target identity.

We test whether these information properties matter in training by intervening on where residual evidence is available and where it is assigned (Figure 3). A geometry-level intervention switches from the non-order-inducing to the order-inducing generator family while holding N, T, nominal density, learner, and optimization protocol fixed. Post-initialization first-entry semantic inactivity

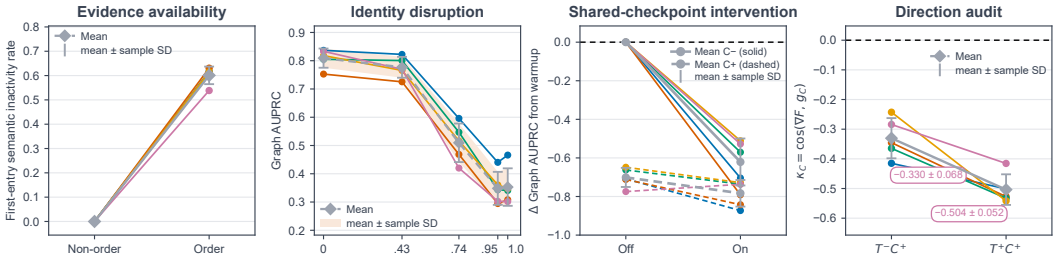


Figure 3: **Calibration and Route probe two distinct failure points in structural verification.** Calibration probes evidence placement and edge–signal correspondence; Route gating and SDA reveal structurally adverse update directions, with $\kappa_C = \cos(\nabla F, g_C) < 0$ in every tested seed.

among false edges increases by 0.601 on average, with the same direction in all five paired seeds, while final Graph AUPRC decreases from 0.944 to 0.809.

A more targeted correspondence intervention branches five dose arms from the same warmup model, optimizer, and RNG state. At each update, nested disjoint transpositions preserve the injected signal matrix's histogram, mean, variance, and norm while progressively altering its node-pair assignment. Using the preregistered identity-disruption fraction as the dose axis, final Graph AUPRC decreases in every seed, with a mean seed-wise slope of -0.510 ± 0.059. Together, the geometry-level and correspondence interventions support the same calibration account: structural fidelity depends both on whether target-discriminating evidence is available and on whether that evidence is assigned to the correct edges within the tested generators.

5.2 Structural Route Alignment

Route concerns the structural direction induced when the task and constraint objectives update the graph.

We branch all four Route arms from the same high-fidelity METR-LA warmup checkpoint and independently gate the task-to-graph and constraint-to-graph paths. Final Graph AUPRC is 0.996 ± 0.002, 0.375 ± 0.121, 0.295 ± 0.050, and 0.213 ± 0.068 for T^-C^-, T^+C^-, T^-C^+, and T^+C^+, respectively. The corresponding task-route main effect is -0.352 ± 0.093, and the constraint-route main effect is -0.432 ± 0.040, with both effects negative in all five paired seeds. Activating either route therefore yields lower fidelity to the declared target than the shared T^-C^- frozen-checkpoint arm in this setting.

The difference-in-differences is $+0.539 \pm 0.085$, indicating sub-additivity on the AUPRC scale. At the same checkpoints, the task–constraint gradient cosine is only -0.041 ± 0.022. The near-zero cosine therefore does not support task–constraint gradient conflict as the explanation, motivating a claim-relative directional audit instead.

To measure this direction directly, we use the **Structural Direction Audit (SDA)**. Let F be a deterministic, differentiable, target-relative local structural-risk probe, and let g_C denote the raw constraint gradient on the graph parameters. Global structural recovery is assessed separately using Graph AUPRC. We define

$$\kappa_C = \cos(\nabla F, g_C).$$

Because the optimizer follows the descent direction $-g_C$, a negative κ_C means that the corresponding descent step locally increases structural risk to first order. In the preregistered pre-floor window, $\kappa_C = -0.504 \pm 0.052$ for T^+C^+ and -0.330 ± 0.068 for T^-C^+, with every seed negative in both arms. The endpoint degradation observed under route gating is therefore accompanied by a directly measured locally adverse constraint-update direction. Appendix E.4 gives the corresponding one-step bound.

PEMS-BAY reproduces the same Route pattern under the matched protocol: the task-route effect is -0.268 ± 0.028, the constraint-route effect is -0.397 ± 0.027, and the difference-in-differences is $+0.391 \pm 0.054$, with the same paired direction in all five seeds. Its pre-floor SDA coefficient is likewise negative in every seed, with $\kappa_C = -0.541 \pm 0.044$. The Route diagnosis is therefore

reproduced across two traffic networks: the training objectives can update the graph along directions that are unfavorable to the declared structural target. This leaves a sharper recovery question: if the diagnosed adverse component is removed, is local directional correction sufficient to restore global structural fidelity?

6 Recovery Boundary and Constructive Implication [D3]

Removing the diagnosed adverse constraint component improves local structural quantities but does not reliably restore global edge ranking, motivating control of structural drift during continued adaptation rather than post-hoc correction after drift has accumulated.

6.1 Oracle Structural-Route Correction

The A^*-dependent oracle is a mechanism test rather than a deployable training rule. Starting from the same warmup state, **Projected** removes the component of g_C anti-aligned with the deterministic structural-risk gradient ∇F, whereas **Magnitude-matched** preserves the original direction while matching the projected gradient norm, controlling for update magnitude. Shared stochastic streams and faithful replay isolate the directional intervention.

Relative to Magnitude-matched, Projected reduces the deterministic structural-risk probe by 5.321 ± 0.579, with the same favorable direction across all paired runs; Precision, Recall, and Density also improve. Graph AUPRC changes by $+0.186 \pm 0.221$, however, with inconsistent run-level outcomes (Figure 4(a)). Thus, correcting the adverse local component improves local structural quantities without reliably restoring global edge ranking.

Smooth local structural quantities and discrete ranking order need not change together: if no positive–negative edge pair crosses a ranking boundary, a structural probe may improve while Graph AUPRC remains unchanged (Appendix E.5). The graded correction sweep likewise yields only partial global recovery (Appendix F.10). Thus, the diagnosed route is structurally relevant, but correcting it alone does not reliably restore the target ranking.

6.2 Warmup-Anchored Structural Guardrail

The recovery boundary shifts the intervention from post-hoc correction to prevention. We therefore constrain post-warmup graph updates around a previously audited warmup anchor; the guardrail does not itself certify that anchor. Let

$$A_w = A^{\mathrm{eval}}_{\theta_G}(T_w), \qquad D_{\mathrm{anc}}(A, A_w) = \frac{1}{N(N-1)} \sum_{i \neq j} |A_{ij} - A_{w,ij}|,$$

where A_w is the detached deterministic soft adjacency at the warmup branch point. For $A_t = A^{\mathrm{eval}}_{\theta_G}(t)$, the guarded post-warmup objective is

$$\mathcal{J}_t^{\mathrm{Trust}} = \mathcal{L}_{\mathrm{task}} + \lambda \mathcal{L}_c + \lambda_t \big(D_{\mathrm{anc}}(A_t, A_w) - \rho_s \big),$$

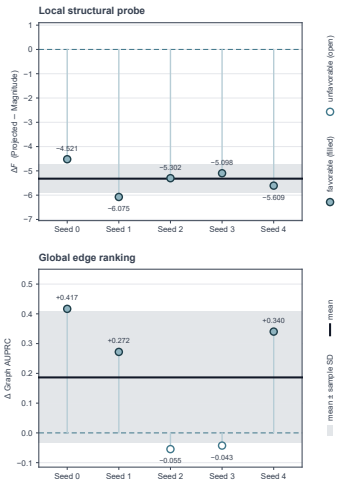
where ρ_s is a per-seed drift radius fixed entirely from pre-branch warmup dynamics. The multiplier λ_t is updated online after each optimizer step; the complete controller update and radius construction are given in Appendix F.13.

The graph learner and all task-model parameters remain trainable under Trust, so the guardrail limits structural movement without freezing adaptation. Its optimization update uses only A_w, ρ_s, and the task and constraint losses. The declared target A^* and structural evaluation metrics are restricted to read-only diagnostics and do not enter the Trust loss, controller, or parameter updates.

We compare three primary non-oracle post-warmup branches: **Original** continues task-and-constraint training without an anchor budget; **Frozen** stops graph-parameter updates while continuing task-model training; and **Trust** applies the warmup-anchor drift budget during continued joint training. Permuted-anchor and support-restricted controls are explicitly labeled and reported separately.

Stage A: real-domain drift mitigation. Relative to Original, Trust improves Graph AUPRC by $+0.751 \pm 0.073$ across the paired runs, with a Trust-minus-Original RMSE difference of $+0.0014 \pm$

(a) Oracle route correction



(b) Warmup-anchor guardrail

Arm	RMSE ↓	AUPRC ↑	Proxy ↓
Stage A: METR-LA			
Original	0.4216 ± 0.0729	0.2128 ± 0.0676	0.0124 ± 0.0073
Frozen	0.4250 ± 0.0700	0.9959 ± 0.0017	0.2999 ± 0.0024
Trust	0.4230 ± 0.0709	0.9640 ± 0.0079	0.2475 ± 0.0082
Stage B: fixed-support weight shift			
Original	0.1581 ± 0.0054	0.3983 ± 0.0419	0.0054 ± 0.0012
Frozen	0.1739 ± 0.0056	0.7014 ± 0.0392	0.0184 ± 0.0058
Trust	0.1653 ± 0.0065	0.6454 ± 0.0356	0.0105 ± 0.0032
Support Oracle†	0.1533 ± 0.0047	1.0000 ± 0.0000	0.0091 ± 0.0012

† The Support Oracle masks the adjacency used for training and evaluation to the true support, thereby restricting graph gradients to that support. It is a non-deployable ceiling, and its Graph AUPRC of 1.0 holds by construction.

Figure 4: **Route correction exposes a recovery boundary, motivating structural preservation.** (a) Projected correction lowers the deterministic structural-risk probe in all five paired runs, but Graph AUPRC improves in only 3/5 runs; the preregistered verdict is therefore `ORACLE_NULL`. (b) The warmup-anchor guardrail limits structural drift on METR-LA and, under the fixed-support weight shift, retains more target fidelity than unanchored adaptation while adapting better than freezing. Values are final-epoch mean ± sample SD over five paired seeds. Lower RMSE/proxy and higher Graph AUPRC are favorable for their measured quantities; the proxy is not fidelity evidence.

0.0048. Trust also ends with a larger Hinge proxy than Original, so its structural retention cannot be attributed to achieving a lower final value of the same surrogate.

Stage B: adaptation under a fixed-support shift. Under the controlled shift, Trust reduces RMSE relative to Frozen by 0.0086 ± 0.0021 and increases Graph AUPRC relative to Original by 0.2471 ± 0.0329, with consistent paired directions across runs. Thus, within this fixed-support setting, Trust achieves higher target fidelity than Original while retaining lower RMSE than Frozen.

These results motivate protecting a previously audited high-fidelity structure during adaptation rather than reconstructing it after drift. The warmup-anchor guardrail operationalizes this strategy without requiring continued access to A^* for post-warmup optimization.

7 Conclusion

Successful constraint optimization and task performance do not, by themselves, establish that a learned graph faithfully realizes its declared structural target. Our experiments reveal two distinct failure points in this evidential link: the constraint interface does not sufficiently distinguish the target from structurally different alternatives, and the realized graph updates move in a target-adverse direction. Correcting the latter improves local structural quantities but does not reliably recover global edge ranking, separating diagnosis from recovery. We further test a preservation strategy in which subsequent adaptation is constrained around a previously audited high-fidelity structure, limiting structural drift without simply freezing the graph. More broadly, our findings indicate that whenever a learned graph is assigned a specific structural interpretation, evidence explicitly tied to the declared target is needed before that interpretation can be treated as verified. This motivates future work to develop training objectives and evaluation interfaces whose reported success can be demonstrably linked to the structures they are intended to realize.

Reproducibility and AI Use

Reproducibility. All reported experimental results derive from archived, frozen, and audited artifacts. The appendix specifies the data, protocols, assumptions, and audit criteria. The anonymous supplement provides a self-contained minimal runnable package for reproducing the principal ex-

periments, analyses, and audits, together with the required configurations, instructions, manifests, and checksums. The complete internal research workspace is retained for provenance but is not included in the submission.

AI Use. Generative AI tools were used to assist with manuscript drafting, language editing, organization, and readability; literature discovery; discussion and refinement of experimental designs; writing portions of the experimental and audit code; auditing and preliminary organization of experimental results; and the formalization and checking of mathematical statements and proof steps.

All cited works were verified by the authors against the original publications and bibliographic records. AI-assisted code was inspected and tested by the authors, and all numerical claims were verified against the underlying experimental outputs. The authors independently checked all mathematical derivations and proofs for correctness and consistency with the implemented objectives and evaluation procedures. The authors made the final decisions regarding the research questions, experimental protocols, interpretation, claims, and manuscript content, reviewed all AI-assisted material, and take full responsibility for the work.

REFERENCES

Humam M. Abdulsahib, Zainab Ahmed Saeed, and Ahmed A. Al-Moadhen. GLIST: Granger-guided dynamic graph learning for directionally interpretable multivariate time series forecasting. *Results in Engineering*, 32:112017, 2026. doi: 10.1016/j.rineng.2026.112017.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Taiyu Ban, Changxin Rong, Xiangyu Wang, Lyuzhou Chen, Xin Wang, Derui Lyu, Qinrui Zhu, and Huanhuan Chen. Differentiable structure learning with ancestral constraints. In *International Conference on Machine Learning*, pp. 2801–2835, 2025.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Linsheng Chen, Bo Zhang, Qingsong Fan, Chen Xia, Hu Fan, Yucan Qiu, and Zhonghua Yu. Anomaly recognition method for cigarette manufacturing process based on graph structure learning. *Computers and Electronics in Agriculture*, 237:110566, 2025. doi: 10.1016/j.compag.2025.110566.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *Advances in Neural Information Processing Systems*, 2020.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. EdNet: A large-scale hierarchical dataset in education. In *Artificial Intelligence in Education*, volume 12164 of *Lecture Notes in Computer Science*, pp. 69–73. Springer, 2020. doi: 10.1007/978-3-030-52240-7_13.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021. doi: 10.1609/aaai.v35i5.16523.

Jean-Antoine Desideri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5–6):313–318, 2012.

Elias Dubbeldam, Reza Mohammadi, Marit Schoonhoven, and S. Ilker Birbil. The role of feature interactions in graph-based tabular deep learning. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=olGaiwoZHZ.

Marc Fischer, Mislav Balunovic, Dana Drachsler-Cohen, Timon Gehr, Ce Zhang, and Martin Vechev. DL2: Training and querying neural networks with logic. In *International Conference on Machine Learning*, pp. 1931–1941, 2019.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *International Conference on Machine Learning*, 2019.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 922–929, 2019. doi: 10.1609/aaai.v33i01.3301922.

Ting Guo, Feng Hou, Yan Pang, Xiaoyun Jia, Zhongwei Wang, and Ruili Wang. Learning and integration of adaptive hybrid graph structures for multivariate time series forecasting. *Information Sciences*, 648:119560, 2023. doi: 10.1016/j.ins.2023.119560.

Nan Jia, Weitao Su, Junrui Xian, Shijia Zou, and Yixue Xia. Adaptive G-UKT: A unified probabilistic framework for knowledge tracing via adaptive graph topology learning and uncertainty-aware gaussian embeddings. *Scientific Reports*, 16:20092, 2026. doi: 10.1038/s41598-026-50711-y.

Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Alessandro Manenti, Daniele Zambon, and Cesare Alippi. Learning latent graph structures and their uncertainty. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42882–42901. PMLR, 2025. URL `https://proceedings.mlr.press/v267/manenti25a.html`.

Ivan Marisca, Cesare Alippi, and Filippo Maria Bianchi. Graph-based forecasting with missing data through spatiotemporal downsampling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34846–34865. PMLR, 2024.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025.

Canberk Ozen, Sławomir Nowaczyk, Prayag Tiwari, and Sepideh Pashami. Assessing the graph structure learning in graph deviation networks. In *Advances in Intelligent Data Analysis XXIII*, volume 15669 of *Lecture Notes in Computer Science*, pp. 97–109. Springer, 2025. doi: 10.1007/978-3-031-91398-3_8.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, 2018.

Chao Shang, Jie Chen, and Jinbo Bi. Discrete graph structure learning for forecasting multiple time series. In *International Conference on Learning Representations*, 2021.

Xun Shi, Kuangrong Hao, Xianyi Zeng, Lei Chen, and Haijian Li. Spatial information bottleneck graph structure learning based multivariate time series prediction for industrial processes. *Journal of Manufacturing Systems*, 85:441–454, 2026. doi: 10.1016/j.jmsy.2026.01.017.

Yuyan Wang, Jie Hu, Fei Teng, Lilan Peng, Shengdong Du, and Tianrui Li. KaTaGCN: Knowledge-augmented and time-aware graph convolutional network for efficient traffic forecasting. *Information Fusion*, 111:102542, 2024. doi: 10.1016/j.inffus.2024.102542.

Dennis Wei, Tian Gao, and Yue Yu. DAGs with no fears: A closer look at continuous optimization for learning bayesian networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Marcel Wienöbst, Leonard Henckel, and Sebastian Weichwald. Embracing discrete search: A reasonable approach to causal structure learning. In *International Conference on Learning Representations*, 2026.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *International Joint Conference on Artificial Intelligence*, 2019.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International Conference on Machine Learning*, 2018.

Kang Xu, Bin Pan, MingXin Zhang, Xuan Zhang, XiaoYu Hou, JingXian Yu, ZhiZhu Lu, Xiao Zeng, and QingQing Jia. Non-end-to-end adaptive graph learning for multi-scale temporal traffic flow prediction. *PLOS ONE*, 20(6):e0322145, 2025. doi: 10.1371/journal.pone.0322145.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163, 2019.

Wei Zhang, Fenghua Zhu, Yisheng Lv, Chang Tan, Wen Liu, Xin Zhang, and Fei-Yue Wang. AdapGL: An adaptive graph learning algorithm for traffic prediction based on spatiotemporal neural networks. *Transportation Research Part C: Emerging Technologies*, 139:103659, 2022a. doi: 10.1016/j.trc.2022.103659.

Weiqi Zhang, Chen Zhang, and Fugee Tsung. GRELEN: Multivariate time series anomaly detection from the perspective of graph relational learning. In *International Joint Conference on Artificial Intelligence*, pp. 2390–2397, 2022b. doi: 10.24963/ijcai.2022/332.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Mingming Gong, Biwei Huang, Kun Zhang, Anton van den Hengel, and Javen Qinfeng Shi. Analytic DAG constraints for differentiable DAG learning. In *International Conference on Learning Representations*, 2025.

Xu Zheng, Farhad Shirani, Tianchun Wang, Wei Cheng, Zhuomin Chen, Haifeng Chen, Hua Wei, and Dongsheng Luo. Towards robust fidelity for evaluating explainability of graph neural networks. In *International Conference on Learning Representations*, 2024.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-Fidelity: A robust framework for faithfulness evaluation of explainable AI. In *International Conference on Learning Representations*, 2025.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Daniel Zügner, François-Xavier Aubet, Victor Garcia Satorras, Tim Januschowski, Stephan Günnemann, and Jan Gasthaus. A study of joint graph inference and forecasting, 2021.

A Datasets, Experimental Units, and Protocol

A.1 Dataset Roles

Table 1: Dataset roles and structural targets. Each row states the scientific role assigned before analysis; a listed dataset is not by itself a claim of cross-domain generality.

Dataset/system	Nodes	Relation semantics	Role in this paper
EdNet KT1	200	Order-inducing prerequisites	Order-domain structural stress test
METR-LA	207	Non-order traffic adjacency	Architecture, form, route, SDA, oracle, and guardrail experiments
PEMS-BAY	325	Non-order traffic adjacency	External dissociation, direct Route/SDA replication, and fail-closed Block-ST graph-identity gate
PEMS08	170	Directed traffic adjacency	External dissociation replication
Synthetic	Fixed configurations	Paired order/non-order generators	Geometry, alignment, and fixed-support adaptation controls
Junyi v7	200	Order-inducing prerequisites	External-validity gate only; task competency blocked

The EdNet KT1 interaction logs originate from the EdNet release (Choi et al., 2020); METR-LA and PEMS-BAY use the benchmark processing established with DCRNN (Li et al., 2018); and PEMS08 uses the PeMSD8 benchmark configuration reported by ASTGCN (Guo et al., 2019). Each structural endpoint uses an adjacency target defined independently of the learned graph. EdNet uses a frozen top-200 concept graph. METR-LA and PEMS-BAY use standard road-network adjacencies. PEMS08 uses a deduplicated 277-edge directed road graph. Synthetic experiments use generator truth. ASSIST is excluded from the strict evidence chain because its skill-index mapping cannot be uniquely reconstructed. Junyi v7 preserves name-level prerequisite relations under audited duplicate-equivalence collapse, but its graph-aware task gate fails.

A.2 Frozen Experimental Matrix

Table 2: Frozen experimental matrix. Formal units use the stated paired seeds and fixed schedules; smoke and superseded runs are excluded.

Unit	Conditions	Seeds	Epochs	Role
P2-A MTGNN	1	5	80	Parameterization replication
P2-A Graph WaveNet	1	5	80	Parameterization replication
P2-B EdNet	1	5	80	Matched-interface order setting
P2-B METR-LA	1	5	80	Matched-interface non-order setting
P2-C Synthetic v3	2	5 paired	80	Geometry intervention and provenance
P3 forms	5	5	80	Enforcement-form supplement
P3b EdNet v3	3	5	80	Order validity, enrichment, and cohort
P3b METR-LA v3	3	5	80	Non-order validity and under-recovery
Binary alignment	2	5 paired	80	Residual placement intervention

Table 2 continued from the previous page

Unit	Conditions	Seeds	Epochs	Role
Alignment dose	5	5 paired	80	Nested-transposition dose response
Route 2×2	4	5 paired	80	Task/constraint route effects
SDA v2	4	5 paired	80	Faithful local direction audit
Oracle route correction	3	5 paired	60 post-warmup	Local adverse-component intervention
PEMS-BAY external v2	2 forms	5 paired	80	Cross-network replication
PEMS08 external v2	2 forms	5 paired	80	Cross-network replication
PEMS-BAY Block-ST R0	3 matched graph arms	5 paired	80	Real-data graph-identity gate; `R0_BLOCKED`, so R1 was not executed
METR-LA parameterization panel	5 learners $\times$ 2 forms	5 paired	80	Four amortized replications + FreeLogit boundary
Synthetic positive 2×2	4	5 paired	80	Calibrated/permuted target manipulation
Long training	2 datasets	5	500	Within-budget persistence
Junyi v7 J1	1 graph-aware encoder	5	20	Task competency blocked; J2 not started
Guardrail Stage A	4 arms	5 paired	60 post-warmup	METR-LA drift mitigation; separately registered Sept. 17 extension
Guardrail Stage B	4 arms	5 paired	60 post-warmup	Fixed-support weight-shift control; separately registered Sept. 17 extension

All units entering formal claims contain seeds $\{0, 1, 2, 3, 4\}$, complete epoch ranges, and unique condition–seed–epoch keys. Junyi is the sole legitimate blocked terminal state: J1 completed and J2 was prohibited by protocol. Smoke tests and superseded results are stored separately and never merged into formal summaries.

A.3 Replication Across Parameterizations and Networks

Table 3: **Replication of the structural-deterioration component of the constraint–structure dissociation.** Values are within-seed final Graph-AUPRC differences, $\Delta = \mathrm{Hinge} - \mathrm{ReLU}$, reported as mean $\pm$ sample SD over five paired seeds. `FreeLogit*` marks the observed parameterization boundary rather than an additional replication.

Scope	System	Δ Graph AUPRC	Direction
Parameterization	Inner-product / METR-LA	-0.767 ± 0.017	5/5 $\downarrow$
Parameterization	MTGNN / METR-LA	-0.622 ± 0.038	5/5 $\downarrow$
Parameterization	Graph WaveNet / METR-LA	-0.410 ± 0.041	5/5 $\downarrow$
Parameterization	EdgeMLP / METR-LA	-0.153 ± 0.057	5/5 $\downarrow$
Network	EdgeMLP / PEMS-BAY	-0.272 ± 0.033	5/5 $\downarrow$
Network	EdgeMLP / PEMS08	-0.384 ± 0.061	5/5 $\downarrow$
Boundary	`FreeLogit*` / METR-LA	0.000 ± 0.000	0/5 $\downarrow$

A.4 Shared Optimization Protocol

Unless an experiment explicitly states otherwise, runs use 80 epochs, Adam with learning rate 10^{-3}, gradient clipping at 1.0, a hard adjacency threshold of 0.5, and a 20-epoch warmup. EdNet uses 10 warmup epochs only in the original matched-interface P2-B run; P3b and long-training controls follow their preregistered schedules. Constraint losses use learned-adjacency weighting with per-active-edge normalization. ReLU-zscore, Hinge, Smooth-scaled, Squared Hinge, and ALM enter at the same post-warmup time within a comparison.

For paired interventions, model initialization, encoder state, graph learner, optimizer state, dataloader order, and RNG stream are shared up to the branch point. Alignment interventions reuse the same signal matrix and transform only its node-pair identity. Route experiments share a full warmup checkpoint and cut selected loss-to-graph gradients while leaving encoder learning active. Oracle probes are deterministic and consume no RNG.

External v2 uses 6,000 traffic time steps and a 150-window held-out set sampled only from a temporally isolated segment. All normalization statistics are computed from the training segment. PEMS-BAY and PEMS08 manifests, code hashes, dataset hashes, and per-cell metadata are identical across compared forms.

A.5 Motivating Literature Audit

Table 4 separates three roles that are easily conflated: using a learned graph for a task, interpreting it structurally, and auditing it against an external reference. The rows are not presented as instances of our exact training schedule or as prior demonstrations of our constraint–structure dissociation.

Table 4: **Claim roles and structural evidence in representative learned-graph work.** "Training access" records whether the external reference enters the learned graph's objective. "Evaluation-only audit" requires a structural comparison whose reference is withheld from the audited post-warmup objectives. A dash means that this evidence is outside the cited study's stated objective.

Work	Structural role or use	External reference and training use	Evaluation-only structural audit	Relation to our setting
GTS (Shang et al., 2021)	Forecasting representation learned from pairwise predictive information	Optional known or neighborhood graph may enter as a regularization prior	–	Prior-guided task-oriented graph learning
Graph WaveNet (Wu et al., 2019)	Forecasting support interpreted as hidden spatial dependency	Predefined and adaptive supports coexist	–	Task-oriented semantic graph
GDN (Deng & Hooi, 2021)	Sensor relations used for anomaly explanation	No external target	Correlation/attention analysis, not target fidelity	Semantic interpretation
Zügner et al. (2021)	Audit of inferred forecasting graphs	Known relations used in graph ablations	Yes	Prediction–structure mismatch
Ozen et al. (2025)	Audit of GDN's graph-learning module	No external target	Module and graph ablations	Component-claim audit
Dubbeldam et al. (2026)	Feature-interaction recovery	Synthetic true graph; imposed in a control	Yes	Prediction–structure mismatch
Shi et al. (2026)	Physical-process topology for interpretable forecasting	Domain graph; not a target loss	Yes	Prediction–fidelity mismatch
CR-GSL (Chen et al., 2025)	Learned relations used for root-cause interpretation	Process graph initialization	–	Semantic interpretation
Adaptive G-UKT (Jia et al., 2026)	Edge weights interpreted as concept-dependency strength	No external target	Visualization, not target fidelity	Semantic interpretation
GLIST (Abdulsahib et al., 2026)	Directional predictive relevance, explicitly non-causal	Granger reference used as a regularizer	Post-hoc diagnostics; reference not held out	Training-guided boundary case
Ours	Fidelity to a declared A^*	Main non-oracle arms: warmup only, then withheld; oracle controls are explicitly labeled	Evaluation after withholding from ordinary post-warmup objectives	Surrogate-certification question

The table establishes neither that our exact warm-start protocol is standard nor that every semantic interpretation entails fidelity to a unique graph. Instead, it documents the narrower motivation: learned graphs are used as predictive representations, semantic objects, and occasionally explicit recovery targets, but these roles impose different evidential obligations. Our controlled protocol isolates the strongest case. Once fidelity to a particular external structure is explicitly claimed, the target is withheld from later objectives so that proxy/task evidence can be distinguished from an evaluation-only post-warmup audit of whether that claim survives adaptation.

B Residual and Form-Aware Diagnostics

For a state tensor $Z \in \mathbb{R}^{B\times N}$ and margin ϵ, the sample-level Hinge violation matrix is

$$V_{ij} = \frac{1}{B}\sum_{b=1}^{B}[Z_{b,j} - Z_{b,i} + \epsilon]_+.$$

For ReLU-zscore, the signal-zero rule is $\mathrm{SMR}_{ij} < \mu_{\mathrm{off}}$, where the mean and standard deviation are computed over off-diagonal entries. Hinge, Squared Hinge, and ALM use whether the corresponding sample-level Hinge violation is zero. Smooth and Smooth-scaled have no strict zero-gradient region. Semantic inactivity uses zero sample-level Hinge violation for every form, so the semantic definition does not change with the optimized surrogate.

For each reference edge, define its empirical violation frequency as

$$q_{ij}(Z) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}[Z_{b,j} > Z_{b,i} + \epsilon].$$

The implemented reference-edge LVR is evaluated only on A^*:

$$\mathrm{LVR}(Z, A^*) = \frac{\sum_{i\neq j} A^*_{ij}\,\mathbf{1}[q_{ij}(Z) > 0.05]}{\sum_{i\neq j} A^*_{ij}}.$$

This diagnostic records the proportion of reference edges whose empirical violation frequency exceeds 5%; it is distinct from the sample-averaged Hinge magnitude V_{ij} used by the training surrogate.

The implemented learned-edge-weighted Hinge objective is

$$\mathcal{L}_c(A^s, Z) = \frac{\sum_{i\neq j} A^s_{ij} V_{ij}}{\sum_{i\neq j} A^s_{ij} + \delta_{\mathrm{num}}}, \qquad \delta_{\mathrm{num}} = 10^{-8}/B.$$

This expression is algebraically identical to the code's batch-summed implementation after dividing numerator and denominator by B. It should be described as a learned-edge-weighted proxy objective, not as physical constraint satisfaction. The numerical stabilizer is negligible for a non-vanishing learned-edge mass, but it means that exact density invariance is an idealization at masses close to zero.

For archived D1–D3 runs, hard-graph density is normalized by N^2; the separately registered guardrail extension instead normalizes over the off-diagonal candidate set, $N(N-1)$. Because diagonal entries are fixed to zero, this convention difference is small, but density is interpreted only within each experimental unit and is not compared numerically across the two packages.

C Transition and First-Entry Protocol

A hard edge birth at epoch t is a transition

$$b_{ij,t} = \mathbf{1}[\hat{A}_{ij,t} = 1,\ \hat{A}_{ij,t-1} = 0,\ i \neq j].$$

The data-generation layer tracks two orthogonal attributes:

- `is_first_entry`: whether edge (i, j) has ever appeared in the run;
- `timing`: `init`, `pre`, or `enforce`, determined solely by epoch and warmup boundaries.

Dynamic-first events satisfy `is_first_entry=True` and `timing in {pre,enforce}`. **Enforce-only** events additionally require `timing=enforce`. Re-entry remains an event but is never counted as a first entry. Candidate-space enrichment is computed against unseen false edges immediately before the transition, preventing future information and re-entry leakage.

For a warmup false-edge cohort, final state is one of three mutually exclusive outcomes: survive, removed, or re-entered. The cohort threshold is fixed at 0.5 and stored in every output row.

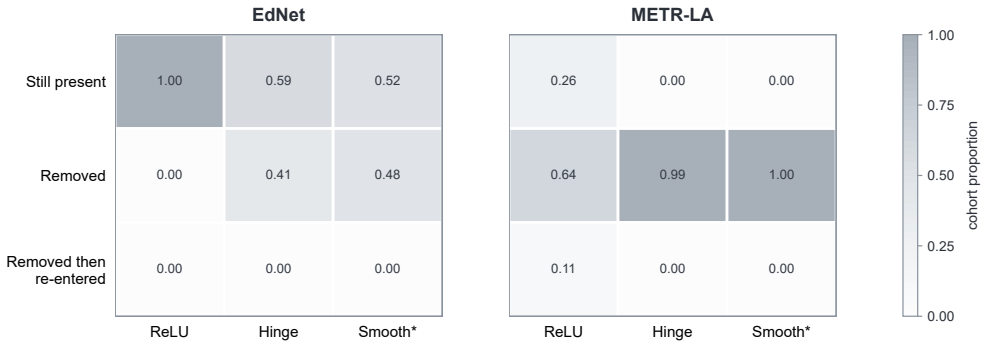


Figure 5: **Outcomes of warmup false-edge cohorts.** Columns denote the ReLU, Hinge, and Smooth* forms; rows report the seed-average proportions of warmup false edges that remain present, are removed, or re-enter by the final epoch. Cohort removal describes the fate of early false edges and does not certify final ranking fidelity to A^*.

D Frozen Numerical Summaries

D.1 Mechanism Ladder

Table 5: First-entry diagnostic summaries across the mechanism ladder. Rank and semantic inactivity are diagnostic observables, not standalone certificates of structural recovery.

Unit	Setting	Dynamic-first rank	Dynamic-first semantic inactivity	Graph AUPRC
P2-A	MTGNN / METR-LA	0.448	–	–
P2-A	GWN / METR-LA	0.516	–	–
P2-B	EdNet / order	0.188	0.566	–
P2-B	METR-LA / non-order	0.508	0.000	0.356
P2-C v3	Order synthetic	0.122	0.601	0.809
P2-C v3	Non-order synthetic	0.490	0.000	0.944

D.2 CONSTRAINT FORMS

Table 6: Constraint-form audit on the repaired P3 protocol. The table documents form behavior; it does not establish form-independence beyond the tested METR-LA setting.

Form	First-entry rank	Final RMSE
ReLU-zscore	0.505	0.428
Hinge	0.500	0.425
Squared Hinge	0.502	0.426
Smooth	0.499	0.423
ALM	0.498	0.425

The repaired ALM run does not collapse; the earlier "third mechanism" interpretation was a software artifact and is excluded.

D.3 PROXY–REALIZATION DECOUPLING

Table 7: Proxy–realization summaries for enforcement-form comparisons. Task metrics are reported descriptively; they are not treated as equivalence tests.

Domain	Form	Constraint proxy/density	Precision	Recall	Task metric
EdNet	ReLU-zscore	LVR 0.498	0.497	1.000	AUC 0.552
EdNet	Hinge	LVR 0.047	0.247	0.747	AUC 0.549
EdNet	Smooth-scaled	LVR 0.065	0.242	0.666	AUC ≈ 0.55
METR-LA	ReLU-zscore	Density 0.006	0.354	0.553	RMSE 0.426
METR-LA	Hinge	Density 0.006	0.041	0.009	RMSE 0.422
METR-LA	Smooth-scaled	Density 0.006	0.031	0.022	RMSE 0.422

For the EdNet headline comparison, final Graph AUPRC is 0.880 under ReLU-zscore and 0.427 under Hinge; the positive-edge prevalence, and hence the expected random-ranking baseline, is 0.021. These ranking quantities audit fidelity to the reference graph and are distinct from LVR.

Smooth-scaled uses $\tau\,\mathrm{softplus}(g/\tau)$ and reproduces the Hinge under-recovery pattern, ruling out the original Smooth form's $1/\tau$ gradient scaling as the sole explanation.

D.4 CALIBRATION, ROUTE, AND RECOVERY

Table 8: Evidence map for Calibration, Route, and the recovery boundary. The interpretation column states the permitted scope of each frozen result.

Evidence unit	Frozen result	Interpretation
Binary alignment	Aligned-minus-permuted AUPRC $+0.0282 \pm 0.0231$, 5/5 positive	Edge-identity placement effect in tested synthetic setting
Alignment dose	Mean slope -0.510 ± 0.059, 5/5 negative	Dose response within nested-transposition family

Table 8 continued from the previous page

Evidence unit	Frozen result	Interpretation
Route 2×2	Task -0.352 ± 0.093; constraint -0.432 ± 0.040; DiD $+0.539 \pm 0.085$	Preregistered route intervention; DiD is sub-additive only
Task–constraint cosine	-0.041 ± 0.022	Near zero; does not support a broad gradient-conflict claim
SDA T^+C^+ pre-floor	$\kappa_C = -0.504$; $\langle \nabla F, g_C \rangle = -0.0053$; $\Delta F = +0.0029$	Faithful local directional diagnosis
PEMS-BAY route/SDA	task -0.268 ± 0.028, constraint -0.397 ± 0.027, DiD $+0.391 \pm 0.054$; $\kappa_C = -0.541 \pm 0.044$	Direct within-traffic replication under Hinge + EdgeMLP
Oracle Projected–Magnitude	AUPRC $+0.186 \pm 0.221$, 3/5; probe and P/R/Density 5/5 improve	`ORACLE_NULL` on primary AUPRC
Graded α-sweep	$R_{\text{recover}} = 0.263$; H1/H2 each 3/5; sham 0.204	Tested constraint-route correction is insufficient for guaranteed recovery
PEMS-BAY Hinge–ReLU	AUPRC -0.272 ± 0.033, 5/5; Hinge proxy $0.306 \to 0.005$ after warmup	Traffic-domain dissociation replication
PEMS08 Hinge–ReLU	AUPRC -0.384 ± 0.061, 5/5; Hinge proxy $0.429 \to 0.002$ after warmup	Traffic-domain dissociation replication
Junyi v7 J1	AUC 0.671 ± 0.004 vs baseline 0.701; 5/5 negative advantage	Task competency blocked

E Conditional Results and Proofs

E.1 Proposition 1: Structural Certificate Characterization

Fix Z. Assume $\mathcal{L}_c^*(Z)$ is finite and consider only $\epsilon > 0$ for which $\mathcal{A}_\epsilon = \{A \in \mathcal{A} : D(A, A^*) \geq \epsilon\}$ is nonempty. Surrogate excess risk is nonnegative by the definition of the infimum.

Forward direction. Assume $\Psi_Z(\epsilon) > 0$ for every admissible ϵ, and let $\mathcal{L}_c(A_n; Z) - \mathcal{L}_c^*(Z) \to 0$. If $D(A_n, A^*) \not\to 0$, then there are $\epsilon_0 > 0$ and a subsequence A_{n_k} such that $D(A_{n_k}, A^*) \geq \epsilon_0$ for every k. The set $\mathcal{A}_{\epsilon_0}$ is therefore nonempty, and

$$\mathcal{L}_c(A_{n_k}; Z) - \mathcal{L}_c^*(Z) \geq \Psi_Z(\epsilon_0) > 0,$$

contradicting convergence of the excess risk to zero. Hence $D(A_n, A^*) \to 0$.

Reverse direction. Assume that surrogate excess converging to zero implies $D(A_n, A^*) \to 0$ for every sequence in $\mathcal{A}$. If $\Psi_Z(\epsilon_0) = 0$ for some admissible ϵ_0, the definition of the infimum gives, for every positive integer n, an $A_n \in \mathcal{A}_{\epsilon_0}$ satisfying

$$0 \leq \mathcal{L}_c(A_n; Z) - \mathcal{L}_c^*(Z) < \frac{1}{n}.$$

The excess risk tends to zero, whereas $D(A_n, A^*) \geq \epsilon_0$ for all n, contradicting the assumed sequential implication. Thus $\Psi_Z(\epsilon) > 0$ for every admissible ϵ. □

Finite-tolerance consequence. Fix an admissible ϵ. If $D(A, A^*) \geq \epsilon$, then $A \in \mathcal{A}_\epsilon$, so

$$\mathcal{L}_c(A; Z) - \mathcal{L}_c^*(Z) \geq \Psi_Z(\epsilon).$$

Taking the contrapositive yields $\mathcal{L}_c(A; Z) - \mathcal{L}_c^*(Z) < \Psi_Z(\epsilon) \Rightarrow D(A, A^*) < \epsilon$. The strict inequality is essential; the equality case requires additional assumptions. □

Claim-relative equivalence classes. No step in the proof requires the zero set of D to be the singleton $\{A^*\}$. If fidelity is defined up to an equivalence class $\mathcal{E}(A^*)$, take $D(A, A^*) = 0$ if and only if $A \in \mathcal{E}(A^*)$. Proposition 1 and the finite-tolerance consequence then certify convergence or finite-tolerance proximity to that claim-relative class. Exact adjacency recovery is the special case $\mathcal{E}(A^*) = \{A^*\}$.

This characterization is conditional on fixed Z, the declared target, the feasible adjacency set, and the chosen structural risk. It supplies a criterion for a certificate; it does not establish that the implemented residual objective satisfies that criterion.

E.2 Lemma 1: Interface-Equivalence Limit

Fix Z, and suppose $\mathcal{L}_c(A; Z) = h_Z(T_Z(A))$, $A^* = A_0 \in \arg\min_{A \in \mathcal{A}} \mathcal{L}_c(A; Z)$, $D(A_1, A_0) > 0$, and $T_Z(A_1) = T_Z(A_0)$.

Proof of Lemma 1. Because the loss factors through T_Z, interface equivalence gives

$$\mathcal{L}_c(A_1; Z) = h_Z(T_Z(A_1)) = h_Z(T_Z(A^*)) = \mathcal{L}_c^*(Z).$$

With $\epsilon_1 = D(A_1, A^*) > 0$, the structurally separated adjacency A_1 has zero excess risk, so $\Psi_Z(\epsilon_1) = 0$. □

Supplementary indistinguishability result. Let q be any structural decision rule observing only the interface statistic, and write the common value as $T_Z(A_0) = T_Z(A_1) = t$. Because q receives the same input t under both structures, its output distribution is identical. Let $p_k = \Pr[q(t) = A_k]$ for $k \in \{0, 1\}$. Since the structures differ, $p_0 + p_1 \leq 1$, and under an equal prior over $\{A_0, A_1\}$,

$$\frac{1}{2} \sum_{k=0}^{1} \Pr[q(t) \neq A_k] = 1 - \frac{p_0 + p_1}{2} \geq \frac{1}{2}.$$

This decision argument also applies to a fixed pair of task/proxy observables after all nuisance quantities are held fixed. Neither the lemma nor the supplementary result asserts that such observational equivalence always occurs, and neither precludes recovery from explicit edge supervision, interventions, or information absent from the interface.

E.3 Lemma 2: Fixed-Mass Residual-Compatible Selection

Fix node states and nonnegative residual costs r_e for candidate edges e. Consider the exactly normalized, fixed-mass idealization with learned edge weights $s_e \in [0, 1]$ and positive total mass $S = \sum_e s_e$. Then

$$\mathcal{L}_c(s) = \frac{\sum_e s_e r_e}{S} \geq \min_e r_e.$$

Equality holds if and only if all positive mass is supported on $\arg\min_e r_e$. With a fixed integer edge budget k, minimizing $\sum_e s_e r_e$ subject to $\sum_e s_e = k$ places unit mass on the k smallest residual costs when the cutoff is unique; with ties, any minimizer may distribute mass within the tied cutoff set. Thus the fixed-mass idealization guarantees residual-compatible selection, not reference-graph recovery. □

The implemented objective in Appendix B replaces S by $S + \delta_{\text{num}}$. For a fixed positive S, this is a positive constant rescaling and has the same ranking of candidate allocations. When S is itself learnable, however, the stabilizer admits an additional mass-shrinkage route and the lemma does not characterize the full optimization dynamics. Jointly changing states, shared parameterization, stochastic sampling, gradient clipping, and Adam momentum require empirical analysis. The matched-density control in Appendix F.7 addresses the corresponding empirical alternative explanation; it does not turn this fixed-mass result into a global theorem.

E.4 Lemma 3: One-Step Structural Route Bound

Let $F(\phi) := F(A^{\text{eval}}_{\theta_G(\phi)}; A^*)$ be an L_F-smooth deterministic structural-risk probe and let $g = g_T + \lambda g_C$ be the graph-parameter gradient. One update $\phi^+ = \phi - \eta g$ satisfies

$$F(\phi^+) \le F(\phi) - \eta\langle \nabla F(\phi), g\rangle + \frac{L_F \eta^2}{2}\|g\|^2.$$

Hence $F(\phi^+) < F(\phi)$ is guaranteed when

$$\langle \nabla F, g\rangle > \frac{L_F \eta}{2}\|g\|^2.$$

For the idealized gradient-descent update, define $\kappa(g) = \cos(\nabla F, g)$. Negative κ indicates that the raw descent direction $-g$ locally increases structural risk to first order. The result is local, depends on the chosen probe, and does not establish long-horizon recovery. □

E.5 Rank-Cell Invariance of Graph AUPRC

Let s_e and s'_e be two score vectors over positive reference edges E^+ and negative candidate edges E^-. Fix the evaluation rule used to group tied scores. Suppose that, for every $(e^+, e^-) \in E^+ \times E^-$, moving from s to s' preserves their cross-class ordering and does not create or break a cross-class tie that changes an evaluation group. Then

$$\text{AUPRC}(s') = \text{AUPRC}(s).$$

In the implementation, reported Graph AUPRC is scikit-learn's `average_precision_score` applied to deterministic soft-adjacency scores and binary reference-edge labels over off-diagonal entries. This average precision depends on the ordered positive/negative label sequence together with its tie grouping. The stated conditions preserve both; permutations within a class do not change the precision–recall sequence. □

Consequently, a smooth structural probe may strictly decrease while Graph AUPRC remains unchanged if no positive–negative pair crosses a ranking boundary. This lemma formalizes a local-to-global distinction; it neither explains every Oracle trajectory nor implies that AUPRC can never improve.

E.6 Metric-Retention Observation

For the warmup anchor A_w, define the realized trust-region violation $v_t = [D_{\text{anc}}(A_t, A_w) - \rho_s]_+$. This analytical positive-part quantity is distinct from the signed, normalized dual-violation diagnostic logged by the implementation. The triangle inequality gives

$$D_{\text{anc}}(A_t, A^*) \le D_{\text{anc}}(A_t, A_w) + D_{\text{anc}}(A_w, A^*) \le \rho_s + v_t + D_{\text{anc}}(A_w, A^*).$$

Thus, if the warmup anchor is already close to A^* under the same normalized L_1 metric and the violation remains small, the guardrail retains an L_1 neighborhood of that target. This is a conditional adjacency-metric observation, not an AUPRC certificate: $D_{\text{anc}}(A_t, A_w)$ does not determine positive–negative edge ordering, and a good or bad anchor can both be retained. The rank-cell lemma above makes the complementary point that metric movement need not change ranking fidelity. Historical warmup artifacts do not include every $D_{\text{anc}}(A_w, A^*)$ value, so we do not retrofit a numerical bound by rerunning or changing the registered endpoint.

E.7 Estimand Distinctions

For edge (i, j), let

$$\tau_{ij} = \min\{t : \hat{A}_{ij,t} = 1,\ \hat{A}_{ij,t-1} = 0\}.$$

First-entry estimands sample each dynamic edge identity at most once. Repeated-transition estimands sample every $0 \to 1$ transition and answer a different question about transition mass. Epoch-balanced summaries weight epochs, not edge identities. The paper uses first-entry estimands for mechanism claims and reports repeated transitions only as supplementary dynamics.

E.8 Scope of the Theoretical Results

These results do not form a universal theorem that constraint-guided GSL must fail. Proposition 1 is a fixed-state certificate characterization; Lemma 1 imposes an interface-only information restriction; Lemma 2 fixes states and idealizes normalized edge mass; Lemma 3 assumes a differentiable probe and local smoothness; and the rank-cell lemma concerns only the ordering that defines AUPRC. Shared parameters, changing states, gradient clipping, Adam momentum, Gumbel sampling, thresholding, and long-horizon path dependence remain empirical components.

F Additional Experimental Details

F.1 Binary Residual-to-Edge Alignment

In the synthetic order setting, aligned and permuted arms share warmup states, model and optimizer states, batches, and RNG streams. The intervention applies a fixed node permutation to the constraint signal only; it does not change the task data, reference graph, signal histogram, mean, variance, or norm. The aligned-minus-permuted Graph-AUPRC contrast is $+0.0282 \pm 0.0231$ with 5/5 positive paired differences. Recall increases by $+0.0010 \pm 0.0022$, while Precision changes by $+0.0455 \pm 0.0139$. We interpret the experiment as a placement effect within the tested synthetic setting, not as an edge-level causal effect.

F.2 Nested-Transposition Dose

The dose family uses nested disjoint transpositions on a 50-node graph. Moved-node fractions are $0, 0.24, 0.48, 0.76, 1.0$, with corresponding pair-identity disruption $0, 0.426, 0.735, 0.946, 1.0$. Each higher dose contains the preceding transpositions, so the intervention path is deterministic and nested. The preregistered seed-level slope is negative in every seed. Because D3 and D4 form a plateau and another permutation family could behave differently, we claim a dose response only for this tested family.

F.3 Task/Constraint Route 2×2

All arms start from a shared METR-LA order-free Hinge warmup checkpoint. T^+ allows task loss to update graph parameters; T^- detaches the graph only on the task route. C^+ allows the constraint objective to update graph parameters; C^- detaches it on the constraint route. The encoder and task head continue to train in all arms. Therefore T^-C^+ is **task-to-graph off**, not a globally "constraint-only" model. Shared RNG streams ensure paired stochastic draws while allowing adjacencies to diverge as logits diverge.

F.4 Structural Direction Audit v2

SDA v1 incorrectly evaluated a stochastic adjacency inside the structural probe and failed faithful replay; it is permanently isolated. SDA v2 evaluates F on a deterministic soft adjacency with zero RNG consumption while retaining training adjacencies for g_T and g_C. Every seed/arm passes a maximum faithful deviation of 10^{-4} or less. We report the full post-warmup window and a preregistered epochs-20–30 pre-floor window. Directional claims use the pre-floor window because later AUPRC saturation reduces local sensitivity; both windows remain archived.

F.5 Oracle Structural-Route Correction

For adverse steps with $\langle \nabla F, g_C \rangle < 0$, Projected applies

$$g_C^{\mathrm{proj}} = g_C - \frac{\langle \nabla F, g_C \rangle}{\|\nabla F\|^2 + \delta} \nabla F.$$

We set $\delta = 10^{-12}$, matching the numerical stabilizer used by the implementation.

Otherwise it leaves g_C unchanged. Magnitude-matched preserves the original g_C direction but scales its norm to match $\|g_C^{\mathrm{proj}}\|$. Unit tests cover negative and nonnegative dots, norm matching, zero gradients, unused parameters, and zero-RNG probes. This experiment is mechanism-closing and non-deployable because it uses A^* at every correction step.

F.6 Alternative-Reference and Null Audits

Stage 0 replays the frozen EdNet and METR-LA soft adjacencies against threshold variants and directed degree-preserving null graphs. METR-LA's null overlap with the original reference is $6.1\% \pm 0.6\%$, and all five seeds exceed the null 95th percentile across the registered variants. EdNet's overlap is $78.2\% \pm 0.7\%$ because its concentrated degree sequence strongly restricts the degree-preserving null space. Consequently, EdNet's absolute AUPRC is partly explained by degree structure; the paper relies on its within-setting LVR–AUPRC dissociation, not on a claim of uniquely recovering prerequisite identity beyond degree.

F.7 Matched-Density Control

The warmup graph was reconstructed in a separate five-seed run and had to reproduce the frozen warmup metrics within 10^{-4} for every seed. Magnitude pruning then matches each Hinge-final edge budget without using residuals. Its AUPRC remains much higher than Hinge-final AUPRC in all seeds. Together with 4–27$\times$ first-entry enrichment in semantically inactive regions, this rejects the explanation that low Hinge AUPRC is merely a mechanical consequence of sparse evaluation.

F.8 Positive Structural Target Control

A synthetic 2×2 crosses target calibration (true A^* versus a permuted target) with task-to-graph route on/off. The calibrated target reaches true-graph AUPRC 1.0 with either route setting, while the permuted target reaches approximately 0.08 against true A^* and 1.0 against the manipulated target. This verifies that the pipeline can recover an explicitly supervised structural target and that evaluation is not inadvertently hard-coded to A^*.

F.9 Junyi v7 Fail-Closed Gate

Junyi users are selected by a stable SHA-256 hash, then assigned to mutually exclusive 70/15/15 splits by an independent hash. The top-200 exercise set is determined using training users only. Two duplicated-name metadata groups are indistinguishable in the interaction log and match exactly on prerequisite sets and all graph/task fields; they differ only in creation date. A preregistered name-level quotient therefore collapses them while preserving both source rows and all dates. The audit enforces

$$E_{\mathrm{post}} = \mathrm{unique}(E_{\mathrm{source}\ \textit{mapped to canonical names}}).$$

The resulting graph has 200 nodes and 208 directed edges. A shared bank of 100 directed degree-preserving null graphs passes in-degree, out-degree, edge-count, self-loop, and duplicate-edge invariants. The graph-aware encoder scores A^* above the null mean in 5/5 seeds, establishing reference sensitivity. Yet test AUC values are $[0.672659, 0.672405, 0.665074, 0.674626, 0.670305]$, with mean 0.671014 and sample SD 0.003656, below the 0.700994 item-frequency baseline in every seed. The runner therefore exits with `TASK_COMPETENCY_BLOCKED`; no J2 full directory exists.

The final audit verdict is `ALL_PASS_BLOCKED_CLOSEOUT`, which means the blocked state itself is complete and reproducible.

F.10 Graded Correction

A graded α-sweep applies fractions $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ of the oracle correction from the same warmup state. It includes a direction-randomized, magnitude-matched sham; a frozen arm; and Direct-A^* as a structural ceiling. Both preregistered slope criteria hold in only 3/5 seeds. The recovery ratio is

$$R_{\text{recover}} = 0.263,$$

compared with 0.204 for the sham, while Direct-A^* reaches AUPRC 1.0. We report the seed heterogeneity rather than replacing the preregistered final-AUPRC endpoint with a post-hoc temporal average.

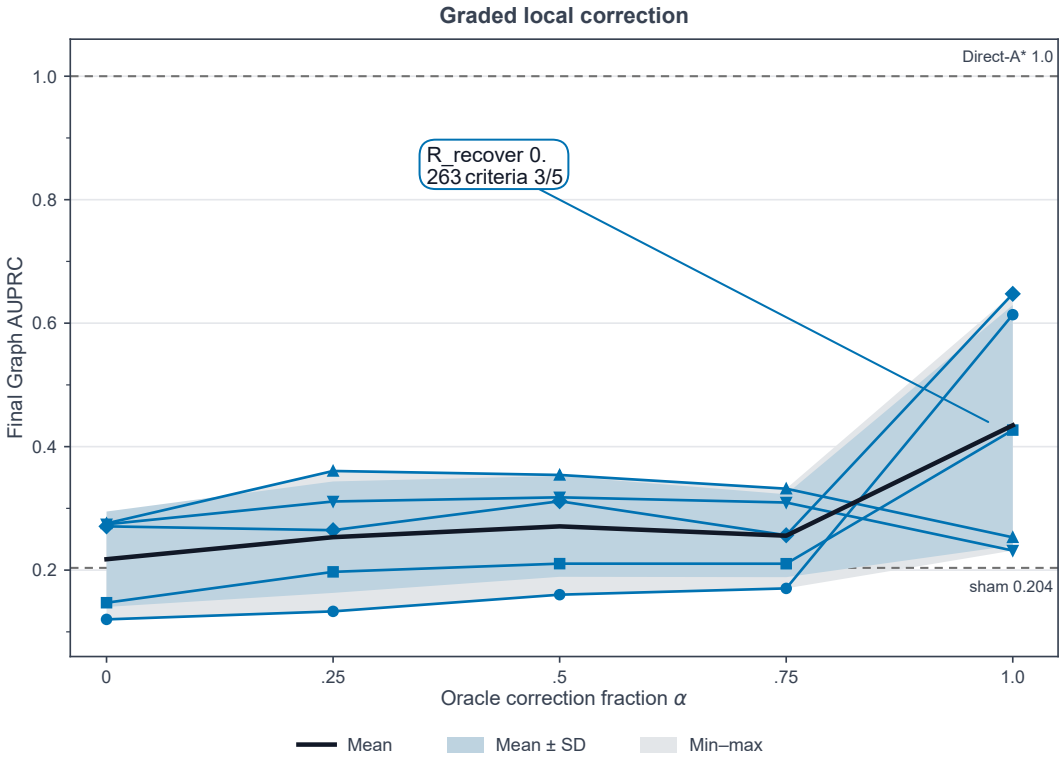


Figure 6: **Graded oracle correction.** Marker-coded trajectories show five paired seeds across the preregistered correction fractions. Lines and bands summarize the seed mean, sample SD, and range; dashed references mark the magnitude-matched sham and Direct-A^* ceiling. The recovery ratio is 0.263, and both direction criteria hold in only 3/5 seeds, showing that stronger local correction does not guarantee global ranking recovery.

F.11 Graph-Dependence Controls

A synthetic graph-required control increases task dependence on graph messages from zero to a preregistered range. The task gate moves from 0/5 at zero graph coupling to 5/5 at coupling at least 0.25, yet AUPRC against A^* decreases from 0.816 to about 0.60 while a decoy graph remains near 0.14. This is a scoped task-necessity control, not evidence of decoy convergence.

On METR-LA, forcing the graph to zero yields only a small median RMSE change (-0.0017, 4/5 negative), indicating a weak task-dependence regime. Conversely, freezing the high-quality warmup

graph isolates a ceiling but does not test post-warmup realization. These controls clarify the role of the learned graph without rescuing the structural claim.

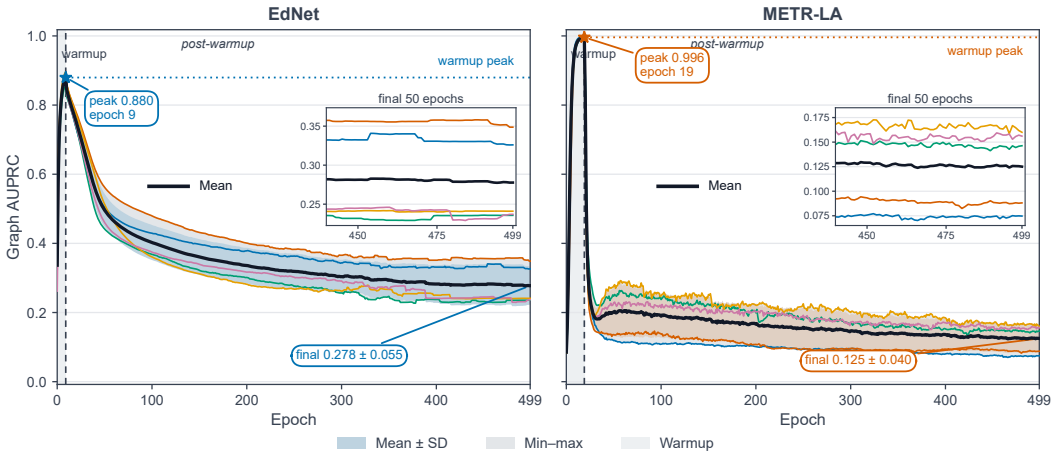


Figure 7: **Persistence under longer training.** Colored trajectories show the five seeds for EdNet and METR-LA across the preregistered 500-epoch budget. Black curves give the seed means, inner bands show mean ± sample SD, outer bands show the seed range, and the shaded interval marks warmup; insets magnify the final 50 epochs. Neither dataset returns to its warmup structural-ranking peak within the tested horizon. This rejects a short-transient explanation over 500 epochs without establishing infinite-time impossibility; bounded late-window slopes can remain slightly positive for individual seeds.

F.12 Fail-Closed Real Block-ST Graph-Identity Gate

We separately registered a real-data gate to test whether a graph-dependent closure experiment was warranted. Adapting the Block-ST missing-input mechanism of Marisca et al. (2024) to PEMS-BAY, we preregistered contiguous faults lasting 12–48 steps. Affected sensors are expanded over up to three road-graph hops with probabilities $0.33, 0.15, 0.05$, and the resulting mask is combined with point missingness; the final fixed mask removes 35.97% of input observations. The target remains the clean next-step speed. Three arms use the same temporal encoder, trainable parameters, windows, mask, initialization rule, optimizer, and five seeds: Graph-On uses the reference road graph; Graph-Off removes cross-node propagation; and the identity control applies one fixed node relabeling to the reference adjacency, preserving its edge count and degree multiset while disrupting road-edge identity. The held-out temporal block is never used for training.

Table 9: **PEMS-BAY Block-ST graph-identity gate.** Held-out RMSE is reported as mean ± sample SD across five paired seeds. Graph-On beats Graph-Off in 5/5 seeds but beats the identity control in only 1/5; the preregistered graph-identity gate is therefore `R0_BLOCKED`.

Arm	Held-out RMSE	Graph-On paired direction
Graph-On A^*	0.5910 ± 0.0419	–
Graph-Off	0.6088 ± 0.0423	5/5 lower
Identity control	0.5820 ± 0.0460	1/5 lower

Graph-On minus Graph-Off is -0.0178 ± 0.0033, corresponding to a 2.92% mean relative reduction. Graph-On minus the identity control is instead $+0.0090 \pm 0.0114$, and Graph-On is lower in only 1/5 seeds. The registered verdict is therefore `R0_BLOCKED`: cross-node propagation helps relative to removing it, but these data do not establish that the reference road-edge identity is task-required. As specified before the full run, no mask ratio, dataset, threshold, or endpoint was changed, and the conditional ReLU/Hinge/ Frozen/Trust closure stage was not executed. This audit narrows rather than removes the real graph-dependence limitation.

F.13 GUARDRAIL STAGE A: METR-LA DRIFT MITIGATION

Both guardrail stages use the same preregistered dual implementation. We set $\lambda_0 = 0$, $\eta_\lambda = 0.01$, and $\lambda_{\max} = 10$. Let $\mathcal{W}_s = \{A^{\text{warm}}_{s,1}, \ldots, A^{\text{warm}}_{s,T_w}\}$ denote the deterministic soft adjacencies recorded after each warmup epoch for seed s, with $A_w = A^{\text{warm}}_{s,T_w}$. Before any post-warmup outcome is observed, the per-seed radius is fixed entirely from these warmup snapshots as

$$\rho_s = \max\Big\{10^{-6},\, \text{median}_{t=T_w-9}^{T_w} D_{\text{anc}}(A^{\text{warm}}_{s,t}, A_w),\, Q_{0.25}(\{D_{\text{anc}}(A, A_w) : A \in \mathcal{W}_s\})\Big\}.$$

After optimizer step k, we recompute the deterministic soft adjacency $A_{k+1} = A^{\text{eval}}_{\theta_{G,k+1}}$ from the updated graph parameters and apply the projected dual update

$$\lambda_{k+1} = \text{clip}_{[0,\lambda_{\max}]}\left[\lambda_k + \eta_\lambda \frac{D_{\text{anc}}(A_{k+1}, A_w) - \rho_s}{\rho_s + 10^{-8}}\right].$$

The multiplier used in the current primal update is therefore updated only after that optimizer step. Neither the radius construction nor this dual controller reads A^* or a structural evaluation metric.

Stage A branches four arms from the same 20-epoch METR-LA Hinge + EdgeMLP warmup checkpoint and runs 60 post-warmup epochs. Original continues joint optimization; Frozen stops graph updates while the encoder and head continue; Trust adds the normalized-L_1 warmup-anchor constraint from Section 6.2; and Permuted Trust uses a fixed cyclically permuted anchor as an identity-sensitive negative control. The Trust optimization path may read A_w, its pre-outcome ρ_s, task loss, and the Hinge surrogate. It does not use A^*, true-edge labels, Graph AUPRC, Precision, or Recall in the Trust loss, dual controller, or parameter updates. Any A^*-dependent quantities computed after warmup are read-only diagnostics and never enter optimization. All arms share the checkpoint, optimizer state, dataloader order, and RNG stream at branching.

At the final epoch, Original, Frozen, and Trust respectively attain Graph AUPRC 0.2128 ± 0.0676, 0.9959 ± 0.0017, and 0.9640 ± 0.0079. Trust minus Original is $+0.7512 \pm 0.0733$, positive in 5/5 paired seeds. Their RMSE values are 0.4216 ± 0.0729, 0.4250 ± 0.0700, and 0.4230 ± 0.0709; Trust minus Original is $+0.0014 \pm 0.0048$, with no equivalence margin or equivalence claim. The corresponding Hinge proxies are 0.0124 ± 0.0073, 0.2999 ± 0.0024, and 0.2475 ± 0.0082. Trust therefore preserves structural ranking while accepting less proxy optimization than Original; the result supports drift mitigation rather than recovery by stronger constraint compliance.

Permuted Trust ends at AUPRC 0.5750 ± 0.0293, but it is not a matched alternative anchor: the permutation makes the warmup graph infeasible at the branch point, and the dual reaches its cap of 10 in every seed. We retain it as a negative control for anchor identity and do not use it to rank guardrail algorithms. Trust's final dual value is 1.168 ± 0.195 and does not saturate.

F.14 GUARDRAIL STAGE B: FIXED-SUPPORT WEIGHT SHIFT

Stage B asks why the guardrail is not simply Frozen-lite. For each seed, a 50-node order-inducing A^* fixes the binary support, while row-normalized propagation weights are deterministically permuted within that support after warmup. The support, edge count, and row mass are invariant. A preregistered world-validity check confirms in 5/5 seeds that a one-step predictor using the post-shift weights outperforms one using the pre-shift weights. Original, Frozen, and Trust have the same post-warmup permissions as in Stage A. Support Oracle alone masks the adjacency used for training and evaluation to A^*, thereby restricting graph gradients to that support; it is explicitly non-deployable.

Final RMSE for Original, Frozen, Trust, and Support Oracle is respectively 0.1581 ± 0.0054, 0.1739 ± 0.0056, 0.1653 ± 0.0065, and 0.1533 ± 0.0047. Frozen minus Trust is $+0.0086 \pm 0.0021$, positive in 5/5 seeds, satisfying the registered adaptation gate. Final Graph AUPRC is 0.3983 ± 0.0419, 0.7014 ± 0.0392, 0.6454 ± 0.0356, and 1.0000 ± 0.0000, respectively. Trust minus Original is $+0.2471 \pm 0.0329$, positive in 5/5, satisfying the registered fidelity gate. Trust remains below Frozen in AUPRC by 0.0560 ± 0.0107 and above Original in RMSE by 0.0072 ± 0.0014, both in 5/5 directions. Hence Trust is non-dominated only among the tested deployable arms at this operating point; we do not infer a Pareto frontier.

The Hinge proxies are 0.0054 ± 0.0012, 0.0184 ± 0.0058, 0.0105 ± 0.0032, and 0.0091 ± 0.0012. Trust's support-weight MAE is 0.1138, between Original (0.0916) and Frozen (0.1340), while its

false-edge mass is 0.1096, between Original (0.0733) and Frozen (0.1258). These diagnostics describe the fixed-support control; they do not extend its conclusion to topology-changing environments. Support Oracle's AUPRC of 1.0 is structural by construction, not evidence for a deployable estimator. Trust's mean final dual value is 0.1233 with activation rate 0.38 ± 0.45; without a fixed- λ comparator, these data do not establish that adaptation of the dual variable is necessary or superior.

G Reproducibility and Audit Index

The accompanying manifests contain machine-resolvable paths, code/data hashes, and checksums. This table is deliberately a human-readable index from each evidence block to its executable entry point and immutable record, rather than a list of filesystem paths.

Table 10: Reproducibility and audit index. Each evidence block is linked to an executable entry point and its immutable audit record.

Evidence block	Executable entry point	Immutable record
P2/P3/P3b	P2/P3 experiment scripts	v3 archive and freeze checksum
Alignment and dose	Alignment intervention scripts	Causal archive and dose manifest
Route 2×2	Route-gating experiment script	Route 2×2 final audit
SDA v2	Structural-direction audit script	SDA v2 final audit
Stage 0 and Oracle	Offline closure and Oracle scripts	Final completion audit
External v2	External replication script	External final audit
v4 parameterization/positive control	v4 extension runner	v4 final audit
v5 route/recovery boundary	v5 extension runners	v5 manifest
v6 long training	Long-training runners	v6 result checksum
Junyi v7	Junyi graph-aware runner	Junyi v7 final audit
Guardrail Stage A	`deployable_repair/run_minimal_repair.py`	`results_deployable_repair/full_v1` audit and checksums
Guardrail Stage B	`deployable_repair/run_adaptation_required.py`	`results_deployable_repair_adaptation/full_v1` audit and checksums
Block-ST graph-identity gate	`blockst_real_closure/run_pipeline.py`	`results_blockst_real_closure/full_v1` audit and checksums
Paper figures	Paper-figure generator	English evidence snapshot

Authoritative status combines the inner process exit code, stage verdict, final audit, and checksums. An outer runner label is never treated as scientific status. Earlier v3 artifacts are anchored by `CHECKSUMS_freeze_20260811.sha256`; later causal, completion, external, and v4–v7 packages are stored in separate directories and never overwrite those anchors.

The Discovery, Calibration, Route, and Oracle evidence chain was frozen on September 5, 2026. Guardrail Stages A and B were separately registered on September 17, 2026 as a constructive extension. Their protocols, result directories, final audits, and checksums are disjoint from the earlier chain; they neither overwrite its artifacts nor retroactively alter its endpoints, gates, or interpretation.

The PEMS-BAY Block-ST gate was registered and archived separately on the same date. Its graph-identity criterion was blocked, so its conditional closure stage did not run and it does not extend the Stage A/B claim.

External v2 did not archive final optimizer and latent states, and some early completion runs did not retain every RNG-state hash. We therefore claim script-level reproducibility and audited result integrity, not bitwise replay for every historical run.

H Claim and Artifact Allocation

Main text: the EdNet dissociation, three-network and four-amortized-learner scope, 500-epoch persistence, Calibration geometry/alignment interventions, Route 2×2, SDA v2, PEMS-BAY route replication, oracle insufficiency, the separately registered Stage A/B guardrail extension, and explicit limitations.

Appendix: form details, first-entry protocol, cohort outcomes, alternative-reference/null audits, matched-density control, positive synthetic target control, complete numerical tables, proofs, and Junyi's fail-closed terminal state.

Permanently excluded from formal claims: ASSIST's unrecoverable skill mapping; old epoch-balanced birth statistics; ALM's pre-fix collapse; SDA v1; external replication v1; Junyi v6 task-aware claims; smoke runs; CASM, CABR, Dynamic-ϵ, structural-recovery guarantees, and universal deployable-repair claims. The allowed constructive claim is narrower: the separately registered warmup-anchor guardrail mitigates drift on METR-LA and yields a tested adaptation–retention compromise in the Stage B fixed-support control.

No additional seeds, datasets, graph learners, or constraint forms were added to the September 5 evidence chain. The September 17 guardrail extension was registered separately before its smoke and full runs and cannot retroactively alter that package. Any future experiment likewise requires a separately registered protocol.

I Scope and Limitations

Limitations. Our certificate question applies when a learned graph is claimed to remain faithful to an externally specified structural target; a purely task-optimal latent adjacency with no such contract need not recover an external reference. The only real order-domain evidence in the strict chain is an EdNet structural stress test with a concentrated reference-degree sequence and weak task dependence; Junyi v7 was correctly blocked. Route interventions are limited to METR-LA and PEMS-BAY under Hinge + EdgeMLP; SDA is a local directional diagnostic in those settings, and the Oracle recovery boundary is limited to METR-LA. Oracle correction requires A^* and is non-deployable. Direction summaries use five paired seeds and do not support large-sample significance or universal claims.

The real Block-ST audit establishes a scoped benefit of cross-node propagation over Graph-Off, but the reference road graph did not outperform the identity-disrupted control consistently. It therefore does not establish a strongly graph-identity-dependent real setting, and no downstream realization or guardrail claim is based on that audit.

The guardrail extension has additional boundaries. Stage A evaluates one real traffic domain, one graph learner, one constraint form, and one registered operating point; it does not show a real-domain advantage over freezing or task equivalence. Stage B is synthetic and changes weights within a fixed true support, not topology. The guardrail retains the warmup anchor and can retain its errors; normalized-L_1 anchor retention is not an AUPRC certificate. No fixed-λ anchoring baseline or ρ sweep was run, so the adaptive dual is neither isolated as necessary nor shown superior to simpler penalties. Support Oracle accesses A^* and is non-deployable. The Stage A/B archives retain scripts, epoch tables, final adjacencies, manifests, audits, and checksums, but not final model and optimizer checkpoints; post-hoc internal diagnostics beyond those artifacts would require a registered rerun.