FALSIFYING OUR OWN MODEL: A PROSPECTIVELY FROZEN PROTOCOL FOR SPATIO-TEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Claims that a model change improves spatio-temporal forecasting often depend on evaluation choices rather than on the change itself. With test predictions fixed, leave-one-comparator-out analyses give six of twelve models at least one new best-in-cell highlight; omitting MetaDG moves Orion from 0/36 to 12/36. Such claims are made one change at a time during development. We apply preregistration principles at this level by **binding each change to a matched control, executable decision rule, and auditable evidence bundle**. A deterministic adjudicator returns a verdict or withholds one when evidence is incomplete; a non-exhaustive ledger records 71 logical events. The three-seed rule crossed in 10.4% of dependent real-seed null splits, whereas bolding the lower mean would have favored the prespecified treatment in 50% of them. Under a normal model, the rule detects a one-standard-deviation effect in 44% of comparisons. Applied to 48 frozen hypotheses on Orion, **the protocol overturned the account we would otherwise have written**. One observational probe and three intervention tests motivated by a diagnosed pattern missed their frozen criteria; the probe lay outside this denominator. All three recipe changes tested on fresh three-seed cohorts crossed the same gate. Orion, although second only to MetaDG in every PEMS08 test cell, ranks fifth of twelve overall. Positive controls recovered two known harms, and the protocol was applied unchanged to STID. Eight of the 48 contracts carry Bitcoin anchors whose block times precede their first results or no-run bounds; five of them do so by more than the two-hour tolerance. Contracts, adjudicators, outcomes, and code are in the supplementary material.

1 INTRODUCTION

A spatio-temporal forecasting paper usually supports a new model with a table of best-in-cell results and a set of ablations, each asserting that one model change helped. These claims are fragile. Which model appears to win depends on the eligible baselines listed. Across leave-one-comparator-out analyses with fixed predictions, six of twelve models gain at least one highlight; omitting MetaDG moves Orion from 0/36 to 12/36 (Figure 1(a)). Under zero effect, bolding the lower three-seed mean favors the prespecified treatment in half of comparisons and names a winner whenever there is no tie (Section 5). Repeated validation search, early test access, seed variation, and rules revised after seeing results shift claims in the same way across machine learning (Zeng et al., 2023; Shao et al., 2025; Melis et al., 2018; Ferrari Dacrema et al., 2019; Bouthillier et al., 2021).

We apply preregistration principles to individual model changes by binding each to a matched control, executable decision rule, and evidence bundle (Section 3). For each comparison, a hash-frozen contract fixes the intervention, control, seeds, metric, rule, and invalidation conditions before its corresponding result; earlier development and test exposures remain disclosed. A deterministic adjudicator returns a verdict or withholds one on missing evidence or failed bindings. A non-exhaustive ledger records 71 logical events, and one test replay follows each closed recipe. The ingredients are known; what is new is binding them to the single change, making every binding checkable by a third party, and measuring how the resulting rule behaves.

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.

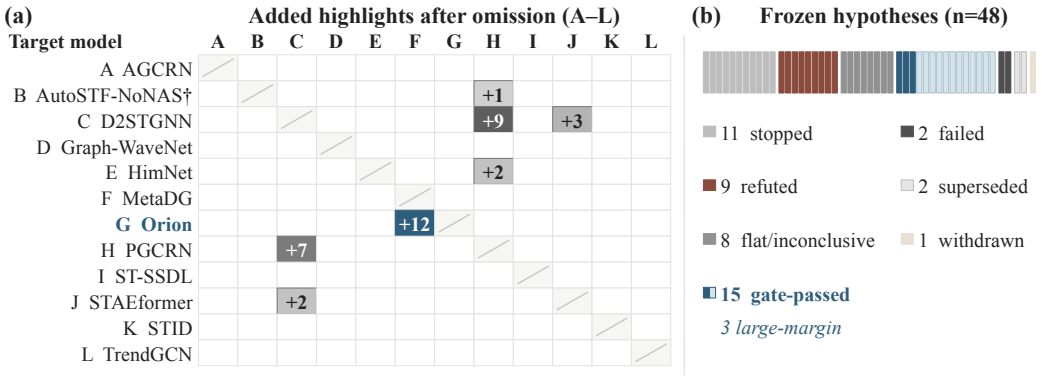

Figure 1: **Comparator sensitivity and prospective self-audit.** (a) Highlights across leave-one-eligible-comparator-out analyses of the fixed test table. Omitting MetaDG changes Orion from 0/36 to 12/36 without prediction changes. Blank cells denote zero added highlights. Diagonal slashes denote self-omissions. † denotes the variant without the official neural architecture search (NAS) weights. (b) Outcomes of the 48 frozen hypotheses: 15 are gate-passed (the registered direction met its frozen rule), and a post hoc replay labels three arms large-margin. Eight contracts have Bitcoin block-header times before their first results or no-run bounds; five precede them by more than the two-hour tolerance. Section 3 defines chronology and scope.

We applied the protocol to the case with the strongest incentive to overclaim: our own model, Orion. Orion is competitive. It is second only to MetaDG in all 12 PEMS08 test cells, which is exactly why one omitted comparator could have credited it with 12 wins. Its development history contains diagnosis-driven repairs and dataset-specific recipes, the kind of changes a paper would normally report as gains. A target-conditioned underprediction pattern, **contraction bias**, motivated one observational probe and three intervention tests; none met its frozen action criterion. The registered weight-decay arm also stayed below its gain criterion. Registry boundaries appear in Appendix D. All three recipe changes tested on fresh seeds, PEMS08 block dropout at $p = 0.35$, the METR-LA graph-rate ratio, and PEMS-BAY graph activation, crossed the same gate again (Appendix Table 27). In the complete 12-model test ledger (Table 1), Orion leads 0 of 36 cells. It ranks fifth (5.528), while PGCRN ranks first (2.583) (Rao et al., 2025). Two positive controls recovered prespecified PEMS08 harms, and six STID comparisons under the unchanged protocol gave four gate-passed, one flat, and one reversed outcome (Appendix M.4). The audit thus reverses the account we would otherwise have written, while keeping every negative result on record.

These results support four contributions:

- **A per-change, executable evidence protocol.** Hash-frozen contracts, a fail-closed adjudicator, and an exposure ledger fix each model change's decision before its result; scripts and documented review exceptions remain separate (Section 3).
- **Measured behavior of the decision rule.** The three-seed rule crossed in 10.4% of dependent real-seed null splits (10.6% simulated), whereas bolding the lower mean would have favored the prespecified treatment in 50% of them. The protocol also recovered two prespecified PEMS08 harms as positive controls; Section 5 reports the rule's statistical operating characteristics.
- **A frozen audit of our own model, and a reproduction ledger.** The audit makes three corrections and one scope clarification to our proposed summary claims, records that the targeted repairs missed their frozen criteria, and carries over unchanged to STID. In our baseline ledger, 120 of 124 comparable mean absolute error (MAE) and root mean squared error (RMSE) cells lie within 5% of their source-reported values (Sections 6 and 7.1).
- **Orion, an openly audited forecaster.** Orion is competitive but not the strongest model: second only to MetaDG in every PEMS08 test cell and fifth of twelve overall. We release it with a documented intervention record that includes negative results.

2 Related Work

Existing work constrains adjacent units of evaluation. Benchmark suites and reassessments expose sensitivity to splits, tuning, leakage, and baseline choice (Zeng et al., 2023; Martin et al., 2026; Shao et al., 2025; Qiu et al., 2024; Zhao et al., 2026; Melis et al., 2018; Ferrari Dacrema et al., 2019; Kapoor & Narayanan, 2023). Checklists preserve reporting and artifact obligations (Reuel et al., 2024; Pineau et al., 2021; Gundersen & Kjensmo, 2018), while multi-seed studies characterize uncertainty (Henderson et al., 2018; Bouthillier et al., 2021; Agarwal et al., 2021). Preregistration commits study plans or submissions (Nosek et al., 2018; Hofman et al., 2023; Thodoroff et al., 2022; Meyer et al., 2026). Adaptive-data-analysis work studies holdout and test reuse (Dwork et al., 2015; Recht et al., 2019; Roelofs et al., 2019), while learner-comparison work studies seed counts and tests (Colas et al., 2018; Benavoli et al., 2017; Dror et al., 2018; 2019). Recent fail-closed proposals commit claim-level scores or tracks, but not one model intervention with a named matched control (Fernández-Barrios et al., 2026; Mitra, 2026).

Our protocol operates at that intervention level. A pre-result contract fixes one control, one change, seeds, metrics, decision and invalidation rules, and permissible downstream use. Deterministic adjudication emits a complete evidence bundle or stops. This composition turns known practices into a prospective, executable audit of one model intervention. Appendices A and B review traffic forecasters, reproducibility studies, the closest proposals, and the methodological boundary.

3 The Protocol

3.1 A claim becomes an adjudicated evidence bundle

The protocol makes each intervention a fixed, checkable object. Each contract records one hypothesis; several displayed comparisons can map to it. Let $C = (D_{\text{split}}, M_c, \delta, S, m, R, F)$ denote the data and split, matched control, sole change, seeds, primary metric, decision table, and invalidation conditions. The registry also retains an explicitly joint graph and width contrast, outside the scope of single-component attribution. The output links a decision to runs, metrics, checks, and exposure receipts. These counts describe different objects. The registry contains 48 hypotheses. Figure 5 has 48 rows (44 mapped, 4 outside the registry). The ledger contains 45 eligible family-level decisions, of which 12 crossed the numerical gate, and 15 final gate-passed hypotheses. "**Gate-passed**" means the registered direction held. It encodes a gain, confirmed harm (e.g., deterministic identity shrinkage), or calibration, so the labels are not 15 benefits. A post hoc replay covered nine treatment arms and called 3 "**large-margin**" below a model-based 0.002 zero-effect crossing probability; this descriptive cutoff changes no verdict.

Algorithm 1: From claim to decision and evidence bundle
Input: claim h and frozen contract C.
1. Bind h to the contract digest and verify result chronology.
2. Locate matched control and intervention runs for every seed on the contracted split.
3. Run the prespecified zero-effect or construction-equivalence check.
4. Read per-seed primary and guard metrics from the actual runs.
5. Stop without a scientific verdict if a seed, equivalence check, or binding is missing.
6. Apply the decision table, including guard failures, without reinterpretation.
7. Audit comparator coverage and selection; stop on violations and disclose test exposure.
8. Return the decision and its contract, metrics, checks, hashes, and receipts.

Script outputs and reviewed rulings are recorded separately: for one dropout contract, the unchanged script returns an inconclusive status, whereas the documented field-mapping review assigns a flat ruling.

3.2 Where the components caught errors

The components catch distinct spurious gains (Figure 2). Their triggers and contract-to-figure mapping appear in Appendices C, D, and E.

Figure 2(b–e,h) maps equivalence, the numerical gate, comparator coverage, test closure, and exposure. Equivalence checks caught implementation drift: the p-norm replay matched all validation

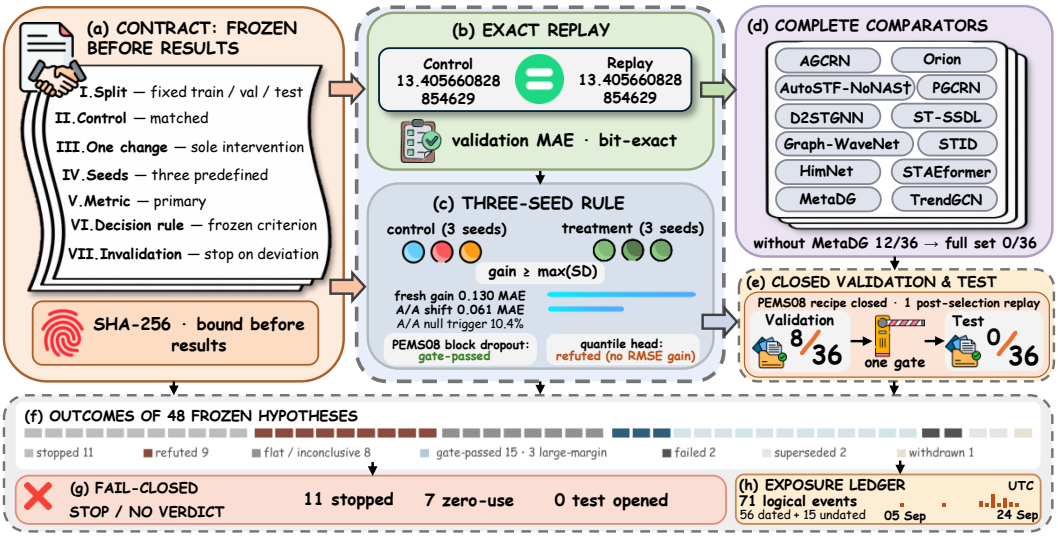

Figure 2: **Prospectively frozen intervention protocol and observed audit trail.** The fresh gain (0.130 validation MAE) and same-configuration control-versus-control (A/A) shift (0.061 validation MAE) share one MAE scale. The 10.4% A/A trigger fraction comes from dependent three-versus-three partitions of 18 real PEMS08 seeds under one configuration; Section 5 gives its statistical interpretation. Panel (e) contrasts 8/36 validation leads among 11 fully covered models with 0/36 test leads among 12 models after earlier test exposure; Section 7.1 states the boundary. The registry has 48 hypotheses in 7 nonempty outcome classes. At final closure, 15 hypotheses were gate-passed. The post hoc replay covered nine treatment arms and labeled 3 large-margin (Section 3). Stopped denotes withholding a confirmatory verdict after a prespecified guard fails. The exposure ledger records 71 known logical events; it is not an exhaustive access history.

errors bit for bit, while the dropout study matched initial-model hashes and parameter counts. For lower-is-better metrics, the three-seed numerical gate is

$$\Delta = \overline{m}_c - \overline{m}_a, \quad s_{\max} = \max\{s_c, s_a\}, \quad \Delta \geq s_{\max}. \tag{1}$$

Here s is sample standard deviation; Section 5 gives the margin's scope. The matched quantile-head comparison returned "no RMSE gain" (Appendix E.1).

Comparator and selection checks expose gains created by evaluation. Completing the opponents changed Orion-leading cells from 12/36 without MetaDG to 0/36 with the complete set without changing predictions. After validation fixes each dataset configuration, the rule permits one prespecified PEMS08 combination and one post-selection test replay even if worse.

Earlier test inspection informed later hypotheses, although validation decisions adopted every current recipe component. The **exposure ledger** (Appendix F) records this prior access; the limitations paragraph states its interpretive boundary.

3.3 Freeze scope, guarantees, and transfer

Prospectively hash-frozen means rules and hashes predate results; public study registration lies outside this claim. Bitcoin header times can differ from true mining times by up to about two hours, so an anchor is decisive only when it precedes the first result by more than that tolerance. Eight of the 48 contracts carry Bitcoin anchors whose block times precede their first results or no-run bounds; five of them do so by more than the two-hour tolerance. The decisive anchors include the fresh-seed PEMS08 dropout replication, our strongest positive result, whose contract was anchored about 10.7 hours before its first result. The remaining chronology categories and record defects are reported in Appendix M.1. The evidence tiers are hash integrity, local chronology, and external anchoring subject to this tolerance.

For complete contracts, the procedure yields a reproducible rule-to-evidence trail and stops on failed bindings. Deterministic scripts adjudicate, and statistical limits are in Section 5.

The same schema tested three STID embeddings. Removing the spatial-identity or time-of-day embedding increased validation MAE on both datasets; removing the day-of-week embedding was flat on PEMS08 and reduced validation MAE on METR-LA (Appendix M.4).

3.4 Cost of stopping

Fail-closed execution stopped 11 times. These stops comprised 8 contract defects, 1 eligibility failure, and 2 non-evasion guards. Seven record zero validation/test/GPU use. None opened test. Costs include repeated seeds, control replays, negative results, and ledgers (Appendix E.1).

4 Case Study Setup

4.1 Orion

Orion (Figure 3) maps a 12-step traffic window $X \in \mathbb{R}^{B\times 12\times N\times 1}$ to a direct 12-step forecast $\widehat{Y} \in \mathbb{R}^{B\times 12\times N\times 1}$. A node-local temporal encoder combines the signal and validity mask with calendar and node-identity features. Temporal blocks and a readout produce a width-W state. A directed influence module learns sample-dependent gates and lags for candidate source–target edges. When enabled, lagged messages correct each future step. Graph-off zeroes the correction.

Audited settings act on these components. W sets state width, and block dropout masks the 64-dimensional node-identity component with probability p per forward call. The graph-rate ratio and activation epoch control graph training. The selected PEMS08 configuration uses $W = 128$ and $p = 0.35$, with graph off throughout. It uses learning rate 0.005 and zero weight decay. METR-LA and PEMS-BAY use $W = 64$ and $p = 0.10$. Both turn the graph on at epoch 31 with a graph learning rate 0.1 times the local rate. No configuration uses input instance normalization. Section 6 reports the audit. Tensor shapes and implementation details appear in Appendix G.

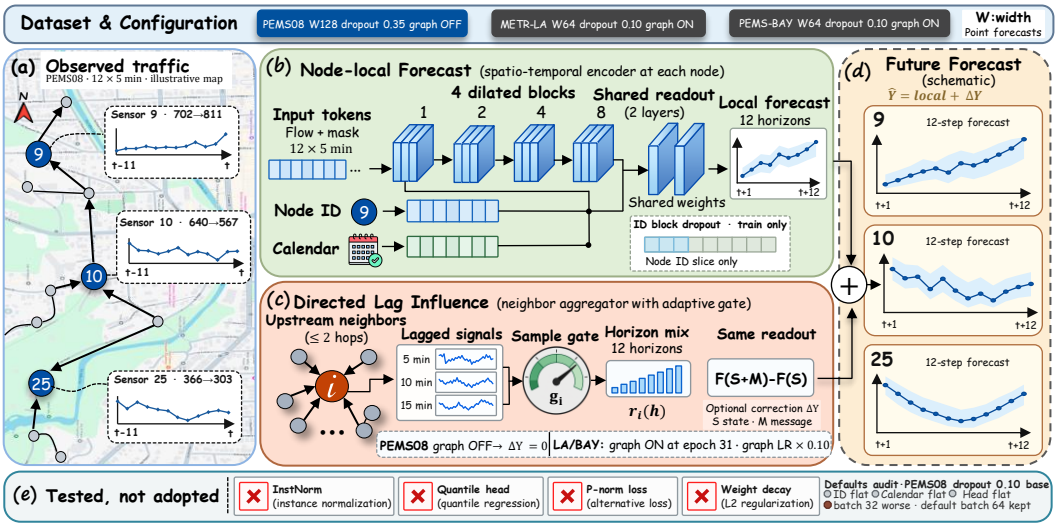

Figure 3: **Orion's input-to-forecast path.** Five regions show observed traffic, node-local forecast, directed lag influence, future forecast, and tested-but-not-adopted choices. PEMS08 histories have 12 real points. The map and point-forecast curves are schematic. Here W denotes width, dropout denotes the node-ID block dropout probability, S denotes local state, and M denotes a directed message. Panel (c) abbreviates the gated correction $\Delta Y = I_{\text{graph}}\kappa_{rh}[F_h(S+M) - F_h(S)]$, whose complete definition appears in Equation (3). ID, calendar, and head settings were flat.

4.2 Evaluation setting

We compare Orion on PEMS08, METR-LA, and PEMS-BAY with 11 reproduced models: AGCRN, AutoSTF-NoNAS†, D2STGNN, Graph-WaveNet, HimNet, MetaDG, PGCRN, ST-SSDL, STAE-former, STID, and TrendGCN. PEMS08 records five-minute traffic flow, whereas METR-LA and

PEMS-BAY record five-minute speed. Each task uses 12 input steps to predict 12 steps. WAPE is weighted absolute percentage error.

Under the frozen eligibility rule, an opponent enters the complete ledger only when its named implementation has accepted three-seed, all-horizon results on the shared data and split, with matching scaling and valid-observation masks. Validation chooses checkpoints and dataset settings before one post-selection test replay. We pool original-scale predictions to compute MAE, RMSE, and WAPE at horizons 3, 6, and 12 and over the 12-step average. TESTAM has only a transferred PEMS08 result without direct PEMS08 training and is excluded from the main table. AutoSTF-NoNAS† lacks the official NAS architecture weights, while PGCRN qualifies through accepted runs but lacks a verifiable source value for reproduction comparison. Training budgets and optimization settings differ.

5 How the Gate Behaves

For one comparison under equal-variance normal/empirical-residual nulls, the three-seed $1 \times s_{\max}$ gate triggers in 10.6%/9.5%. It detects one/two-SD shifts at 44%/82%. Across 25 independent nulls (hypothetical repetitions of this single-comparison generator), at least one crosses with probability 94%. Shared controls require separate analysis (Appendix Table 9).

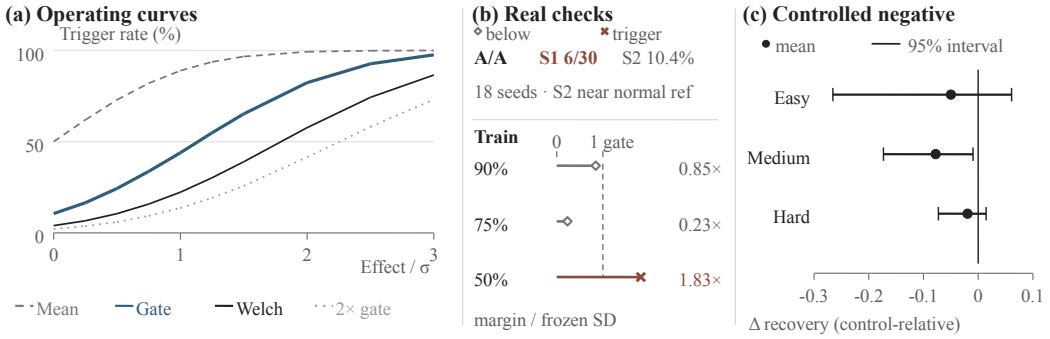

Figure 4: **Gate behavior and controlled negative case.** (a) Trigger rate against effect in common standard deviations (SD, σ). The gate uses the larger arm SD. Mean-bolding (simply bolding the lower mean) and a one-sided Welch test at $\alpha = 0.05$ are reference rules, not competing forecasters. (b) In the same-configuration A/A analysis, S1 is 6/30 ordered comparisons among six disjoint three-seed blocks. S2 is 10.4% dependent partitions of the same 18 seeds. The lower part of panel (b) shows training-fraction degradation with 90%, 75%, and 50% of training data retained. The dashed line marks the $1\times$ gate. (c) Susceptible–exposed–infectious–recovered (SEIR) Δ recovery relative to the best control: tier means and 95% intervals. Appendix Figure 8 and Table 10 give details.

The exhaustive real-seed partitions match the model-based null rate (10.4% versus 10.6%), whereas six fixed three-seed blocks crossed in 6/30 ordered comparisons (20%) because three blocks had small within-block SDs. Control and treatment share seed identifiers, while the gate uses the larger unpaired arm SD; under this design, the zero-effect crossing rate therefore depends on same-seed correlation (Appendix Table 7). Where seeds are shared, paired intervals agree with the gate in 11 of 15 comparable decisions (Appendix M.6). Under this null, bolding the lower mean favored the prespecified treatment in 50% of comparisons and named a winner whenever there was no tie. Semi-synthetic one- and two-SD shifts were detected at 41% and 81% (Appendix E.2).

Because decisions reuse controls, a global-null replay yields 4.80 expected crossings and $\Pr[K \geq 12] = 0.065$ over 45 decisions. Post hoc, Benjamini–Hochberg (BH) correction flags 8 at $q = 0.10$ (6 at $q = 0.05$), while Holm flags 4 at $\alpha = 0.05$. The model-based tail probabilities are descriptive for this adaptive analysis and change no label (Appendices E.3 and M.2).

In the training-fraction control, only 50% retention crossed and the response was nonmonotone. In a separately frozen synthetic SEIR experiment, Orion's mean Δ against the best control was negative in all three tiers; none passed, and no positive recovery control was included (Figure 4(b,c); Appendix Table 10).

Positive controls. Under one frozen contract, two contrasts reversed PEMS08 normalization choices whose harm had been measured earlier. The reversals lowered validation MAE by 0.927 (12.5 $s_{\max}$) and 0.175 (1.97 $s_{\max}$); both gains fell inside bands prespecified around the earlier estimates, so the rule recovered both known harms (Appendix M.3).

Statistical interpretation and scope. The $1 \times s_{\max}$ gate is a pre-result, loss-uncalibrated rule, not a test, confidence statement, or effect size. With three normal seeds, paired-difference SD has about 52% relative variation. Moving from $1\times$ to $2 \times s_{\max}$ lowers both the null rate (10.6% to 2.1%) and one-/two-SD detection (43.9%/82.3% to 13.6%/41.5%). A gate-passed label admits one frozen comparison, not efficacy. Its value is pre-outcome fixation and an auditable record; **fail-closed** refers only to stopping on invalid evidence (Appendix M.2).

6 What the Protocol Found in Orion

We combine three fresh-seed replays with large-margin observations under the scope in Section 5.

6.1 Targeted repairs did not meet their frozen action criteria

The target-conditioned underprediction pattern motivated one observational probe, the METR-LA latent-separability probe, and three intervention tests: PEMS08 p-norm masked loss, earlier METR-LA graph activation, and the PEMS08 five-quantile head (Figure 5; Appendix H). None met its frozen action criterion, and the pattern serves only to motivate these tests, leaving forecast miscalibration and correctability unresolved. The probe and both instance-normalization arms were frozen outside the 48-contract registry (Appendix D). Weight decay was a separate registered component check and remained below its frozen gain criterion.

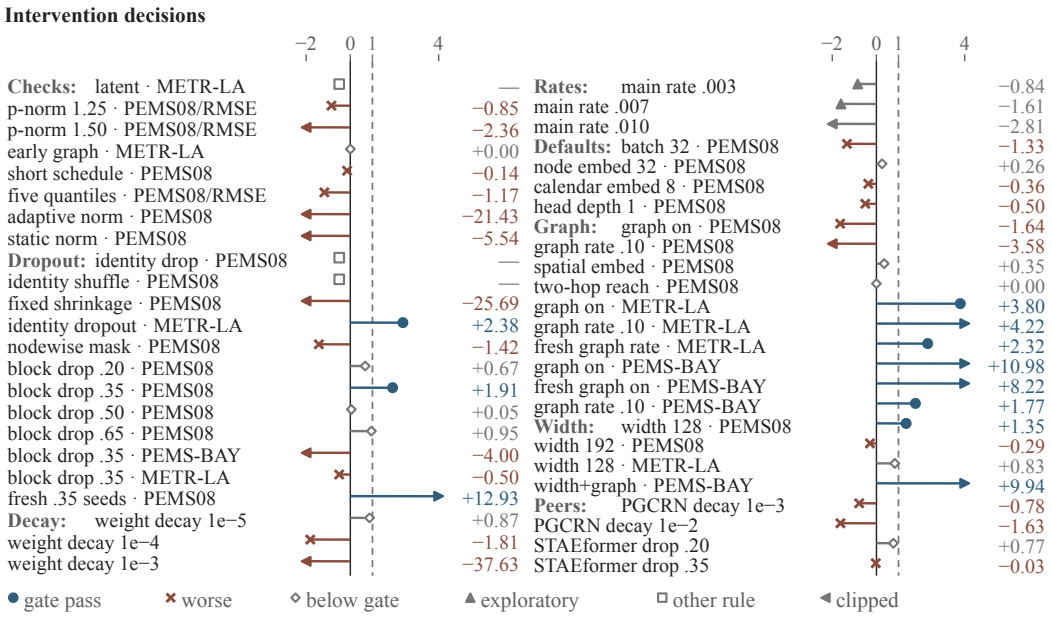

Figure 5: **Intervention audit with 48 displays and 45 eligible family-replay decisions.** Bars show $m = (\bar{x}_{\text{control}} - \bar{x}_{\text{arm}})/s_{\max}$ from zero. The dashed line is the gate. Arrows clip outside $[-2, 4]$, while printed margins do not. Squares indicate other frozen rules, and triangles indicate exploratory main-rate arms. The Checks group contains four contraction-bias-motivated tests and auxiliary checks. "Fixed shrinkage" is deterministic node-identity scaling, unrelated to contraction bias. Appendix Tables 20 and 27 report omitted numerical crossings. Remaining values and contract mapping appear in Appendices E.1 and D.1.

Within the Checks group, both higher-order masked-loss arms and the five-quantile head worsened RMSE; both normalization arms worsened MAE. A compressed schedule and earlier graph activation missed their criteria.

6.2 WHERE ARE THE GAINS?

All three recipe changes tested on fresh seeds passed again: PEMS08 block dropout, METR-LA graph-rate ratio, and PEMS-BAY graph activation; no others were, including PEMS08 width 128 (Figure 5; Appendix Table 27). On a fresh three-seed cohort (seeds 3410–3412), the METR-LA graph-rate contrast gained 0.01505 MAE against a threshold of 0.00648, and PEMS-BAY graph activation gained 0.01903 MAE against 0.00231. The post hoc large-margin label covers PEMS-BAY graph activation, its joint graph–width arm, and the METR-LA graph-rate ratio.

Fresh PEMS08 dropout improved validation MAE by 0.130, versus a 0.061 same-configuration A/A shift. Six-seed pooling gave a descriptive 0.109. A separately contracted 18-seed-per-arm study, frozen before its 12 new runs but designed after the three-seed results, estimated a 0.107 gain with an unpaired Welch analysis (95% interval $[0.087, 0.127]$), and every three-per-arm subset crossed the gate. The selected interval is in-sample and unadjusted for selection (Appendix M.5).

The dropout effect is dataset-specific: 0.65 lowered mean MAE but missed the extra-improvement rule against 0.35 ($0.953\, s_{\max}$), while 0.35 worsened PEMS-BAY and was flat on METR-LA. The joint graph–width arm changed both components and cannot isolate either (Appendix E.5).

Weight-decay, main-rate, default, and limited PGCRN/STAEformer audits yielded no admitted gain; halving batch size crossed the worse-direction rule. The recipe is unchanged (Appendix E.6).

6.3 THREE CORRECTIONS AND ONE SCOPE CLARIFICATION

Beyond component selection, one audit chain produced three corrections and one scope clarification of nonindependent summary claims under the scope in Section 5. First, "Orion is the most stable model" becomes "its 11.160% worst gap ranks 5th of 12 models," behind HimNet (4.950%), PGCRN (5.656%), STAEformer (5.661%), and AutoSTF-NoNAS† (10.1%). Second, the PEMS08-scoped statement "MetaDG is its only stronger PEMS08 opponent" remains true but is incomplete across datasets: MetaDG leads all 12 PEMS08 cells, D2STGNN and STAEformer lead nine and three METR-LA cells, and PGCRN leads all 12 PEMS-BAY cells. On PEMS08, MetaDG's reproduced MAE and RMSE averages differ from its source values by only 0.11% and about 0.25%. Third, the audit replaces the older average-cell stability summary with the complete ledger's 11.160% worst gap. Fourth, "Orion leads 12/36 cells" becomes "with the complete comparator set, Orion leads 0/36; all 12 apparent leads came from omitting MetaDG." Scope details and 9 formal `REFUTED` labels appear in Appendix Table 4.

7 BEYOND ORION: THE COMPLETE LEDGER

7.1 THE COMPLETE TEST LEDGER

Across 36 correlated test cells and 12 fully covered models, Orion leads none, ranks 5th (5.528), and has an 11.160% worst gap; PGCRN ranks first at 2.583 (Table 1). The selection-validation ledger has 11 complete models: Orion leads 8/36 validation cells versus 0/36 on test (Appendix F). Across the six accounting choices in Appendix Table 36(a–f), Orion ranks fifth or sixth and never higher; by dataset it is second on PEMS08. Dropping AutoSTF-NoNAS† or MetaDG moves its mean rank to 4.86, still fifth; counting only MAE and RMSE moves it to 5.71, sixth. The dataset-level positions (second on PEMS08, eighth on METR-LA, sixth on PEMS-BAY) show where the model is strong and where it is not.

Across 10,000 descriptive three-seed resamples, Orion always led 0/36, while PGCRN ranked first or tied first in 98.7% (Appendix Table 23).

Comparator-set sensitivity. Across twelve leave-one-comparator-out analyses, at least one omission creates highlights for 6/12 models (resampling range 5–7). Omitting MetaDG gives Orion 12/36 without changing predictions (Appendix Table 23).

Table 1: **Complete test ledger under frozen eligibility (Section 4.2), not a field-wide ranking.** Entries are 12-step mean MAE/RMSE/WAPE (%); rank and worst gap use all 36 displayed cells. Leads give full set/maximum after one omission. Training budgets and search effort differ. Shading only locates Orion; no cell is bolded as best (Section 5).

Model	PEMS08	METR-LA	PEMS-BAY	Mean rank	Worst gap (%)	Leads (full / max omit 1)
PGCRN	13.74/23.05/5.9	2.91/5.99/5.0	1.50/3.45/2.4	2.58	5.7	12/19
HimNet	13.56/23.28/5.9	2.93/6.06/5.1	1.52/3.54/2.4	3.89	5.0	0/2
D2STGNN	14.43/24.02/6.2	2.89/5.91/5.0	1.52/3.52/2.4	4.22	13.2	9/18
STAEformer	13.51/23.30/5.8	2.94/5.97/5.1	1.54/3.56/2.5	4.64	5.7	3/5
Orion	13.23/22.98/5.7	3.07/6.42/5.3	1.54/3.59/2.5	5.53	11.2	0/12
AutoSTF-NoNAS†	14.23/23.25/6.1	3.00/6.20/5.2	1.53/3.53/2.5	5.56	10.1	0/1
MetaDG	13.03/22.47/5.6	3.71/9.24/6.4	1.53/3.56/2.4	5.72	63.7	12/12
ST-SSDL	14.86/23.90/6.4	2.92/6.03/5.1	1.60/3.65/2.6	7.89	18.1	0/0
STID	14.20/23.39/6.1	3.12/6.48/5.4	1.56/3.60/2.5	8.08	12.1	0/0
Graph-WaveNet	14.83/23.71/6.4	3.05/6.17/5.3	1.57/3.63/2.5	8.33	15.7	0/0
TrendGCN	15.13/24.15/6.5	3.13/6.35/5.4	1.58/3.65/2.5	10.00	17.2	0/0
AGCRN	15.91/25.14/6.9	3.19/6.45/5.5	1.64/3.75/2.6	11.56	24.4	0/0

All 12 models cover 36/36 test cells; WAPE is displayed as a percent. †AutoSTF-NoNAS omits the official NAS architecture weights. Tables 22 and 23 give full horizons and seed stability.

Reported and reproduced values. Of 186 source-reported cells, 124 MAE/RMSE cells are comparable: 120 reproduced values differ from their source-reported values by at most 5% (relative to the source value), and the 4 weaker cases are ST-SSDL. Mean absolute percentage error (MAPE) is incompatible, PGCRN lacks a verifiable source value, and protocol equivalence is not implied; implementation bias remains (Table 24; Appendix J).

A second model. Under the unchanged protocol, removing STID's spatial-identity or time-of-day embedding raised validation MAE on both datasets (by 8.0 to $76.2\,s_{\max}$), while removing the day-of-week embedding was flat on PEMS08 and lowered it on METR-LA (Appendix M.4).

Limits of the audit. Earlier test exposure and greater search limit interpretation of Orion's test results. Two positive controls cover only large/moderate effects; smaller-effect detection remains simulated. The STID transfer covers one model family, and each embedding removal changes parameter count. WAPE duplicates MAE rankings within each dataset–horizon pair. GPU/person-hour costs remain unquantified.

8 Recommendations, Limitations, and Conclusion

With fixed predictions, omitting MetaDG moved Orion from 0/36 to 12/36. Our frozen executable protocol measured the decision rule's behavior under the null; one observational probe and three intervention tests missed their action criteria, while all three recipe changes tested on fresh seeds passed again. This audit makes three corrections and one scope clarification to our summary claims, and Orion ranks fifth overall. The same record keeps every result that did not favor us, which is what gives readers reason to trust the ones that did. The executable per-change protocol, measured null behavior, self-audit and reproduction ledger, and openly audited Orion are the four contributions.

Recommendations. For forecasting papers that report ablations, we recommend four practices: (i) report best-in-cell counts together with leave-one-comparator-out sensitivity, as in Figure 1(a); (ii) freeze each ablation's control, seeds, metric, and decision rule before running it, and record script outputs separately from reviewed rulings; (iii) measure the decision rule's null crossing and detection rates on one's own seeds before relying on it (Section 5); and (iv) publish a test-exposure ledger next to the results table, stating which pre-result claims are externally anchored. **Outlook.** The protocol is ready for stronger tests than a self-audit can provide: contracts registered by independent teams, a hidden future test period released only after closure, a prespecified minimum effect size alongside the gate, and measured protocol cost. Each extends the same contract object rather than replacing it.

AI USE STATEMENT

We conceived, designed, and directed this research, and we made all scientific decisions reported in this paper. We used large language model tools (Anthropic Claude Code and OpenAI Codex) in a supporting role: for literature search, feedback on the methodology and experimental design, assistance with implementing and running experiments and analyzing data, and drafting and language editing of the text and figures. We reviewed and verified all AI-assisted material before including it, and no AI tool is listed as an author. We take full responsibility for the content of this paper, including its code, results, and text.

REPRODUCIBILITY STATEMENT

The anonymous supplementary material contains the complete audit trail: for each of the 48 frozen hypotheses, the hash-frozen contract, its SHA-256 sidecar and OpenTimestamps receipt where one was made, the adjudicator script, its saved output, and the acceptance record; per-seed validation and test results for Orion and every reproduced baseline; the test-exposure ledger; and the stopping records of fail-closed attempts. Its `README.md` maps the numbers, tables, and figures in the paper to the files that produce them, and `REPRODUCE.md` gives CPU-only commands that rebuild Table 1 and the data figures from acceptance records and re-run any adjudicator for a byte-level comparison with its saved output, plus entry points for full retraining. Runs record code, configuration, data, environment, and source-bundle hashes; Orion used Python 3.11.10 and PyTorch 2.4.0 with CUDA 12.1 on NVIDIA RTX 3090 GPUs, with main seeds 3407–3409 and fresh or A/A seeds 3410–3424, and positive-control seeds 3425–3427, as declared in the contracts. Baselines ran from public code at pinned revisions; the STID second-model case reused its accepted baseline implementation. The adapters log every deviation from the original recipe (Section 4.2 and Table 24). A public code repository will be released with the camera-ready version.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f514cec81cb148559cf475e7426eed5e-Paper.pdf.

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17804–17815. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ce1aad92b939420fc17005e5461e6f48-Paper.pdf.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017. URL https://jmlr.org/papers/v18/16-305.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL http://jmlr.org/papers/v11/cawley10a.html.

Andrea Cini, Ivan Marisca, Daniele Zambon, and Cesare Alippi. Taming local effects in graph-based spatiotemporal forecasting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing*

Systems, volume 36, pp. 55375–55393. Curran Associates, Inc., 2023. doi: 10.52202/075280-2417. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ad58c61c71efd5436134a3ecc87da6ea-Paper-Conference.pdf.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? Statistical power analysis in deep reinforcement learning experiments, 2018. URL https://arxiv.org/abs/1806.08295.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL http://jmlr.org/papers/v7/demsar06a.html.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1224. URL https://doi.org/10.18653/v1/D19-1224.

Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 631–641. ACM, 2024. doi: 10.1145/3637528.3671961. URL https://doi.org/10.1145/3637528.3671961.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance—how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266/.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 307–317. ACM, 2025. doi: 10.1145/3690624.3709177. URL https://doi.org/10.1145/3690624.3709177.

Roberto Fernández-Barrios, Iker Pastor-López, Asier González-Santocildes, and Pablo García Bringas. Conditional validity of quantum event classifiers under collider systematics and quantum estimation uncertainty, 2026. URL https://arxiv.org/abs/2609.02781.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 101–109. ACM, 2019. doi: 10.1145/3298689.3347058. URL https://doi.org/10.1145/3298689.3347058.

Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, Yuxin Ma, and Xuan Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 3998–4006, 2024. doi: 10.24963/ijcai.2024/442.

Haotian Gao, Zheng Dong, Jiawei Yong, Shintaro Fukushima, Kenjiro Taura, and Renhe Jiang. How different from the past? Spatio-temporal time series forecasting with self-supervised deviation learning. In *Advances in Neural Information Processing Systems*, volume 38 of *NeurIPS 2025*, pp. 146647–146675. Curran Associates, Inc., 2025. doi: 10.52202/085713-4905. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d7b351608d824a4680344a02b180a947-Paper-Conference.pdf.

Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Association for the Advancement of Artificial Intelligence (AAAI), 2018. doi: 10.1609/aaai.v32i1.11503. URL https://doi.org/10.1609/aaai.v32i1.11503.

Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. BigST: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5):1081–1090, 2024. doi: 10.14778/3641204.3641217. URL https://doi.org/10.14778/3641204.3641217.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Association for the Advancement of Artificial Intelligence (AAAI), 2018. doi: 10.1609/aaai.v32i1.11694. URL https://doi.org/10.1609/aaai.v32i1.11694.

Jake M. Hofman, Angelos Chatzimparmpas, Amit Sharma, Duncan J. Watts, and Jessica Hullman. Pre-registration for predictive modeling, 2023. URL https://arxiv.org/abs/2311.18807.

Juyong Jiang, Binqing Wu, Ling Chen, Kai Zhang, and Sunghun Kim. Enhancing the robustness via adversarial learning and joint spatial-temporal embeddings in traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 987–996. ACM, 2023. doi: 10.1145/3583780.3614868. URL https://doi.org/10.1145/3583780.3614868.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804. URL https://doi.org/10.1016/j.patter.2023.100804.

Hyunwook Lee and Sungahn Ko. TESTAM: A time-enhanced spatio-temporal attention model with mixture of experts. In *International Conference on Learning Representations*, volume 2024, pp. 3011–3029, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0cb310ed8121549488fea8e8c2056096-Paper-Conference.pdf.

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. URL https://openreview.net/forum?id=SJiHXGWAZ.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *Queue*, 17(1):45–77, 2019. doi: 10.1145/3317287.3328534. URL https://doi.org/10.1145/3317287.3328534.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pp. 4125–4129. ACM, 2023a. doi: 10.1145/3583780.3615160. URL https://doi.org/10.1145/3583780.3615160.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. LargeST: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 36 of *NeurIPS 2023*, pp. 75354–75371. Curran Associates, Inc., 2023b. doi: 10.52202/075280-3293. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ee57cd73a76bd927ffca3dda1dc3b9d4-Paper-Datasets_and_Benchmarks.pdf. Datasets and Benchmarks Track.

Tengfei Lyu, Weijia Zhang, Jinliang Deng, and Hao Liu. AutoSTF: Decoupled neural architecture search for cost-effective automated spatio-temporal forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 985–996. ACM, 2025. doi: 10.1145/3690624.3709323. URL https://doi.org/10.1145/3690624.3709323.

Kenneth Martin, Simon Heilig, Asja Fischer, Michel F. C. Haddad, Adam M. Sykulski, and Moshe Eliasof. A critical audit of spatiotemporal forecasting benchmark datasets and baselines, 2026. URL https://arxiv.org/abs/2608.20980.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByJHuTgA-.

Marcel Meyer, Sascha Kaltenpoth, Henrik Albers, Kevin Zalipski, and Oliver Müller. TS-Arena – a live forecast pre-registration platform. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9558–9568. ACM, 2026. doi: 10.1145/3770855.3817515. URL https://doi.org/10.1145/3770855.3817515.

Arkid Mitra. The MODA general attribute suite: A four-track evaluation benchmark for fashion attribute extraction, 2026. URL https://arxiv.org/abs/2609.13279.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, volume 12370 of *Lecture Notes in Computer Science*, pp. 681–699. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58595-2_41. URL https://doi.org/10.1007/978-3-030-58595-2_41.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114. URL https://doi.org/10.1073/pnas.1708274114.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL http://jmlr.org/papers/v22/20-303.html.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. *Proceedings of the VLDB Endowment*, 17(9):2363–2377, 2024. doi: 10.14778/3665844.3665863. URL https://doi.org/10.14778/3665844.3665863.

Xuan Rao, Shuo Shang, Renhe Jiang, Lisi Chen, and Peng Han. Traffic forecasting with patch-based graph convolutional recurrent network. *GeoInformatica*, 29(3):691–720, 2025. doi: 10.1007/s10707-025-00544-0. URL https://doi.org/10.1007/s10707-025-00544-0.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J. Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems*, volume 37 of *NeurIPS 2024*, pp. 21763–21813. Curran Associates, Inc., 2024. doi: 10.52202/079017-0685. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9175–9185. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/ee39e503b6bedf0c98c388b7e8589aca-Abstract.html.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, pp. 4454–4458. ACM, 2022a. doi: 10.1145/3511808.3557702. URL https://doi.org/10.1145/3511808.3557702.

Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S. Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *Proceedings of the VLDB Endowment*, 15(11):2733–2746, 2022b. doi: 10.14778/3551793.3551827. URL https://doi.org/10.14778/3551793.3551827.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, Gao Cong, Christian S. Jensen, and Xueqi Cheng. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):291–305, 2025. doi: 10.1109/TKDE.2024.3484454. URL https://doi.org/10.1109/TKDE.2024.3484454.

Pierre Thodoroff, Wenyu Li, and Neil D. Lawrence. Benchmarking real-time reinforcement learning. In *NeurIPS 2021 Workshop on Pre-registration in Machine Learning*, volume 181 of *Proceedings of Machine Learning Research*, pp. 26–41. PMLR, 2022. URL https://proceedings.mlr.press/v181/thodoroff22a.html.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019, pp. 1907–1913. International Joint Conferences on Artificial Intelligence Organization, 2019. doi: 10.24963/ijcai.2019/264. URL https://doi.org/10.24963/ijcai.2019/264.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, IJCAI-2018, pp. 3634–3640. International Joint Conferences on Artificial Intelligence Organization, 2018. doi: 10.24963/ijcai.2018/505. URL https://doi.org/10.24963/ijcai.2018/505.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128. Association for the Advancement of Artificial Intelligence (AAAI), 2023. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Zhiyuan Zhao, Juntong Ni, Shangqing Xu, Haoxin Liu, Wei Jin, and B. Aditya Prakash. TimeRecipe: A time-series forecasting recipe via benchmarking module level effectiveness. In *International Conference on Learning Representations*, volume 2026, pp. 150745–150792, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f46ddea413df86832418c5e04e59644f-Paper-Conference.pdf.

Yiqing Zou, Hanning Yuan, Qianyu Yang, Ziqiang Yuan, Shuliang Wang, and Sijie Ruan. Meta dynamic graph for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16584–16592. Association for the Advancement of Artificial Intelligence (AAAI), 2026. doi: 10.1609/aaai.v40i19.38699. URL https://doi.org/10.1609/aaai.v40i19.38699.

Appendix Contents

A Extended Related Work **16**

B Boundary to Adjacent Evaluation Practices **17**

C Protocol component examples **17**

D Intervention registry **17**

D.1 Contract-to-arm display crosswalk 21

E Protocol and Orion adjudication details **23**

E.1 Contract and protocol audit details 23

E.2 Exact gate-behavior replay values 24

E.3 Gate calibration sensitivity and family replay 24

E.4 Decision-rule comparison and controlled negative case 25

E.5 Cross-dataset dropout-strength boundary 26

E.6 Capacity, graph-rate, and dropout response curves 27

F Recorded historical test-exposure ledger **27**

G Orion architecture and implementation **35**

H Detailed Orion intervention evidence **38**

I Audit-boundary ledger **40**

J Reported-versus-Reproduced Result Ledger **42**

K Additional Audit Figures **46**

L Selected Validation Cases and Error Contributions **46**

M Additional evidence added before submission **46**

M.1 Bitcoin anchoring of timestamp receipts 46

M.2 Post hoc multiplicity and paired-seed summaries 47

M.3 A positive control 52

M.4 A second model under the same protocol 53

M.5 Eighteen seeds for the PEMS08 dropout effect 54

M.6 Rank-accounting sensitivity and paired comparisons 55

N Claim-to-Evidence Map **56**

A Extended Related Work

Spatio-temporal forecasting. Traffic forecasting progressed from fixed road graphs such as DCRNN and STGCN to learned or dynamic relations in Graph-WaveNet, AGCRN, and D2STGNN (Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Bai et al., 2020; Shao et al., 2022b). STID and STAEformer foreground identity and embeddings. Later systems study heterogeneity, experts, operator choice, deviation-aware training, generated graphs, scale, and patches (Shao et al., 2022a; Liu et al., 2023a; Dong et al., 2024; Lee & Ko, 2024; Lyu et al., 2025; Gao et al., 2025; Zou et al., 2026; Han et al., 2024; Fang et al., 2025). Patch-based PGCRN has the lowest mean test-cell rank within our frozen eligible roster (Rao et al., 2025). Cini et al. (2023) analyze node embeddings and local effects. Orion uses node-ID conditioning, while its graph gains vary by dataset. This methodological variety makes component-level attribution difficult. Our roster is bounded by the frozen eligibility rule (Section 4.2) rather than by a survey of recent forecasters; self-supervised pretraining frameworks such as STD-MAE (Gao et al., 2024), which also report PEMS08, METR-LA, and PEMS-BAY results, lie outside it.

Benchmarks, reproduction, and reporting. Benchmark reassessments make comparison design central. Linear baselines can rank ahead of examined Transformers, and traffic-specific audits revisit METR-LA and PEMS-BAY baselines (Zeng et al., 2023; Martin et al., 2026). BasicTS+ and TFB standardize time-series evaluation, LargeST expands traffic data, and TimeRecipe studies component effects (Shao et al., 2025; Qiu et al., 2024; Liu et al., 2023b; Zhao et al., 2026). Reassessments in language modeling, metric learning, and recommendation show how tuning and baseline choice revise reported gains (Melis et al., 2018; Musgrave et al., 2020; Ferrari Dacrema et al., 2019). BetterBench evaluates benchmark quality, while Pineau et al. document reproducibility checklists, code policies, and replication efforts (Reuel et al., 2024; Pineau et al., 2021). Scholarship norms, reporting completeness, and leakage analyses provide additional safeguards (Lipton & Steinhardt, 2019; Gundersen & Kjensmo, 2018; Kapoor & Narayanan, 2023). These approaches improve a benchmark or study, but do not specify a fail-closed verdict for every model change.

Preregistration, seeds, and repeated selection. Study-level and predictive-model preregistration record plans (Nosek et al., 2018; Hofman et al., 2023), and prospective machine-learning experiment proposals serve a related role (Thodoroff et al., 2022). Seed and sampling variation affect benchmark conclusions (Henderson et al., 2018; Bouthillier et al., 2021). Few-run uncertainty has also been quantified for reinforcement-learning benchmarks (Agarwal et al., 2021). Search budgets, repeated selection, and multiple comparisons require separate accounting (Dodge et al., 2019; Cawley & Talbot, 2010; Demšar, 2006). TS-Arena accepts forecast submissions during a preregistration window before ground truth exists. Its commitment unit is the prediction submission (Meyer et al., 2026). Our unit is instead one model intervention with a named matched control, a frozen decision rule, and explicit invalidation and downstream-use conditions.

Adaptive-data-analysis work studies validity under repeated interaction with held-out data, while benchmark reassessments document how test-set reuse and distribution shift can erode apparent generalization (Dwork et al., 2015; Recht et al., 2019; Roelofs et al., 2019). Statistical comparisons of learning systems likewise emphasize seed count, paired structure, multiplicity, and uncertainty rather than a single run (Colas et al., 2018; Benavoli et al., 2017; Dror et al., 2018; 2019). Our narrower distinction is procedural: the decision rule is frozen for each individual change and then used to audit our own model. We do not claim that this rule has greater inferential power than a careful paired or Welch analysis. Its contribution is advance specification and an auditable record of whether the declared rule was followed.

Closest fail-closed proposals and empirical boundary. Recent preprints report fail-closed per-claim verdicts for quantum classifiers and hash-committed scoring with negative interventions in a vision benchmark (Fernández-Barrios et al., 2026; Mitra, 2026). Relative to MODA's track-wise prediction commitments, our protocol freezes a single-variable model intervention, records test exposure, and measures the decision rule's operating behavior. We do not claim to invent preregistration, multi-seed evaluation, reproducibility audits, or immutable experiment records. The added object is a mechanically adjudicated claim-evidence bundle that either returns a complete verdict or stops. In our case study, six pre-run STID component comparisons produced four gate-passed, one flat, and one reversed outcome. Two known-harm PEMS08 reversals were recovered on fresh

seeds. Eight of 48 Orion contracts have Bitcoin block-header times before their first results or no-run bounds; five precede them by more than the two-hour tolerance. Restoring MetaDG changes Orion's apparent test-cell leads from 12/36 to 0/36 without changing predictions, and failed or unresolved attempts remain in the record.

B Boundary to Adjacent Evaluation Practices

Table 2 separates the commitment unit, chronology proof, test-exposure accounting, mechanical verdict, comparator sensitivity, and failure-stopping rule reported by each adjacent evaluation practice.

Table 2: Methodological boundary with BasicTS+ (Shao et al., 2025), TFB (Qiu et al., 2024), TimeRecipe (Zhao et al., 2026), BetterBench (Reuel et al., 2024), and public study-level preregistration (Nosek et al., 2018; Thodoroff et al., 2022). NR means the specified per-intervention mechanism is not reported by the cited work, not that it was never used. Chronology remains partial. Eight of 48 registered contracts have Bitcoin block-header times before their first results or no-run bounds, five of them beyond the two-hour tolerance; later or absent anchors cannot witness an earlier freeze. Separate cases submitted to public calendars before their runs recovered two known-harm reversals and adjudicated six STID component comparisons. Their contracts were later anchored in Bitcoin block 968554, before the first P169 result but after the first P170 result (Appendix M). BetterBench's checklist scores benchmark quality, a different verdict target.

Work	**Commitment unit**	**Chronology proof**	**Test exposure accounting**	**Mechanical verdict**	**Comparator sensitivity**	**Failure stopping**
BasicTS+	benchmark	NR	NR	NR	NR	NR
TFB	benchmark	NR	NR	NR	NR	NR
TimeRecipe	module	NR	NR	NR	NR	NR
BetterBench	benchmark	NR	NR	checklist	NR	NR
Public prereg.	study plan	public plan	NR	NR	NR	NR
This study	intervention	partial	✓	✓	✓	✓

C Protocol component examples

Table 3 links each frozen protocol component to the failure mode it guards against and the concrete Orion event that exercised that guard.

Table 3: **Each protocol component has a concrete trigger in this study.** Decisions apply only to the contracted intervention and data scope.

Component	**Failure addressed**	**Orion trigger**
Frozen contract	Mutable success rule	The p-norm loss arms were bound to one digest and decision table.
Zero-effect check	Implementation drift	The $p = 1$ loss path exactly replayed control validation error.
Mechanical gate	Selective interpretation	The five-quantile head was labeled "no RMSE gain."
Completeness audit	Missing opponents	Restoring eligible opponents removed apparent Orion-leading cells.
Fail-closed check	Incomplete evidence	A lead–lag qualification stopped with zero validation, test, and GPU consumption.
Selection and exposure	Test reuse	The validation-selected graph-rate checkpoints retained a disclosed test opening.

D Intervention registry

The accepted snapshot contains 48 registered hypotheses. The outcome column uses gate-passed for an arm that passed its frozen rule. This wording does not strengthen the frozen ruling. Stopped de-

notes withholding a confirmatory verdict after a prespecified guard fails; seven of the eleven stopped cases explicitly record zero validation, test, and GPU use, and none opened the test split. Registry IDs match the contract filenames in the supplementary material. One frozen calendar-dropout hypothesis was withdrawn before execution, and the fresh-seed replication passed both frozen arm-level decisions. The post-selection test replay is a protocol step, not an additional hypothesis. It is recorded in the arm-level audit crosswalk below. Table 4 lists the accepted registry.

Table 4: Accepted registry of 48 frozen hypotheses, listing each contract ID, claim, and terminal outcome.

Registry ID	Description	Recorded outcome
O4-NODE-CVAR-20260918	A node-tail conditional value at risk (CVaR) training objective should improve validation performance by emphasizing high-error nodes.	Refuted
P097-V3-GRADGEOM-ATTRIBUTION-20260918	The earlier target-alignment arm gain can be attributed to target alignment rather than gradient-clipping geometry.	Stopped
P098-GAIN-DECOMPOSITION-20260918	The earlier target-alignment arm gain can be decomposed into separable gradient-clipping and Adam-epsilon effects.	Failed
P099-ARMS-DIRECT-20260918	At least one direct clip-only or epsilon-only arm should reproduce the earlier target-alignment arm gain against the matched graph-off control.	Stopped
P100-ARMS-DIRECT-HARDENED-20260918	A hardened direct-arm test should isolate whether clipping or Adam epsilon causes the earlier target-alignment arm gain.	Stopped
P101-SCDL-QUALGATE-20260918	Lead-lag relations between physically adjacent nodes should vary with traffic state beyond a calibrated null.	Stopped
P102-NSNM-20260918	Node-scale mismatch should measurably confound learned graph-message strength.	Stopped
P103-DIRPRIOR-20260918	Adjacent-node dependencies should exhibit detectable directionality that can improve a directed physical prior.	Stopped
P104-DIRPRIOR-MIN-20260918	A minimally specified directionality test should detect asymmetric adjacent-node dependence beyond its null.	Failed
P105-GAPLOC-20260918	Orion's validation gap to matched MetaDG should concentrate on a temporally stable subset of nodes.	Stopped
P106-GAPLOC-BOUND-20260918	The matched Orion-MetaDG validation gap should be concentrated rather than uniform across nodes.	Gate-passed
P107-INSTNORM-20260918	Instance normalization should reduce the backbone validation gap independently of additional parameters.	Stopped
P108-INSTNORM-BOUND-20260918	A fully bound instance-normalization intervention should reduce the backbone validation gap independently of extra parameters.	Stopped
P109-INSTNORM-CONTSUPP-20260918	A continuous-support qualification should validate instance normalization before testing its prediction gain.	Stopped
P110-INSTNORM-RNGFROZEN-20260918	A random-number-generator (RNG)-frozen qualification should validate instance normalization before testing its prediction gain.	Stopped
P112-IDENTITY-SHUFFLE-CONTROL-20260919	A node-identity shuffle control should distinguish identity-specific benefit from generic regularization.	Refuted
P113-GRAPH-CONTRIBUTION-20260919	Enabling graph propagation under identity damping should improve validation MAE over the matched graph-off model.	Refuted
P114-STOCHASTICITY-ISOLATION-20260919	The identity-dropout gain should require stochastic masking rather than deterministic scaling alone.	Gate-passed

Registry ID	Description	Recorded outcome (continued)
P115-METRLA-DROPOUT-CONTROL-20260919	Identity block dropout should improve METR-LA validation MAE under a matched configuration.	Gate-passed
P117-GRAPH-ON-METRLA-20260919	Enabling graph propagation should improve METR-LA validation MAE over the matched graph-off model.	Gate-passed
P118-GRAPH-ON-PEMSBAY-20260920	Enabling graph propagation should improve PEMS-BAY validation MAE over the matched graph-off model.	Gate-passed
P120-DROPOUT-STRENGTH-METRLA-20260920	Increasing identity block-dropout strength should improve METR-LA validation MAE.	Superseded
P121-BUDGET-MATCHED-SCHEDULE-PEMS08-20260920	A budget-matched schedule change should improve PEMS08 validation MAE.	Flat or inconclusive
P124-CAPACITY-SCALING-PEMS08-20260920	Increasing hidden capacity should improve PEMS08 validation MAE.	Gate-passed
P125-DISABLED-SWITCHES-PEMS08-20260921	Enabling the disabled adaptive spatial embedding (ASE) or two-hop graph switches should improve PEMS08 validation MAE.	Refuted
P126-GRAPH-TRAINING-BUDGET-METRLA-20260921	Graph performance should be limited by graph learning rate or graph-active training budget.	Gate-passed
P129-BEST-ORION-METRLA-20260921	Increasing hidden_dim from 64 to 128 should improve METR-LA validation MAE when graph propagation is enabled.	Flat or inconclusive
P129C-CAPACITY-W128-PEMSBAY-20260922	Registered as a width hypothesis, but the frozen gate compared graph-off W64 with graph-on W128. The recorded label therefore applies to this joint contrast and does not identify a width-only effect.	Gate-passed
P131-PNORM-LOSS-PEMS08-20260922	Raising the masked-error norm from p=1 to p in 1.25, 1.50 should reduce PEMS08 validation RMSE without an unacceptable MAE cost.	Refuted
P132-CAPACITY-W192-PEMS08-20260922	Increasing hidden_dim from 128 to 192 should further improve PEMS08 validation MAE.	Flat or inconclusive
P133-GRAPH-LR-RETEST-PEMS08-20260922	This test asks whether graph propagation remains harmful to PEMS08 validation MAE after graph_lr_ratio is raised from 0.02 to 0.1.	Gate-passed
P134-CALENDAR-BLOCK-DROPOUT-PEMS08-20260922	Adding calendar-slice block dropout on top of identity-slice block dropout should improve PEMS08 validation MAE through an interaction effect.	Withdrawn before execution
P136-BLOCK-DROPOUT-GRANULARITY-PEMS08-20260922	Replacing one global identity-dropout draw with independent per-node draws at the same expected masking rate should improve PEMS08 validation MAE.	Refuted
P137-GRAPH-LR-PEMSBAY-20260922	Raising graph_lr_ratio from 0.02 to 0.1 should improve PEMS-BAY validation MAE as it did on METR-LA.	Gate-passed
P138-GRAPH-LR-UPPER-BOUND-METRLA-20260922	Graph propagation should remain undertrained at graph_lr_ratio=0.1, so at least one higher ratio should improve METR-LA validation MAE.	Flat or inconclusive
P140-QUANTILE-HEAD-PEMS08-20260923	A five-quantile output head with a frozen mean reconstruction should reduce PEMS08 validation RMSE without excessive MAE degradation.	Refuted
P141-DROPOUT-STRENGTH-PEMS08-20260923	Increasing block_dropout_p above 0.1 should improve PEMS08 validation MAE.	Gate-passed

Registry ID	Description	Recorded outcome (continued)
P142-DROPOUT-STRENGTH-PEMSBAY-20260923	Increasing block_dropout_p from 0.1 to 0.2 should improve PEMS-BAY validation MAE.	Superseded
P145-DROPOUT-STRENGTH-PEMSBAY-20260923	Increasing block_dropout_p from 0.1 to 0.35 should improve PEMS-BAY validation MAE.	Refuted
P147-DROPOUT-STRENGTH-METRLA-20260923	Increasing block_dropout_p from 0.1 to 0.35 should improve METR-LA validation MAE.	Flat or inconclusive
P148-DROPOUT-SWEEP-COMPLETION-PEMS08-20260923	Increasing block_dropout_p beyond 0.35 to 0.50 or 0.65 should reveal whether PEMS08 validation MAE continues improving or reaches a boundary.	Flat or inconclusive
P149-UNDERREG-AND-LR-PEMS08-20260923	H1 tests whether nonzero weight decay improves PEMS08 validation MAE, while H2 explores whether the main learning rate is undertuned.	Flat or inconclusive
P156-P157-GATE-CALIBRATION-PEMS08-20260923	The same-configuration check records whether the frozen three-seed decision gate triggers under a real A/A comparison, while the nested-data check tests whether nested training-data reductions produce the expected degradation response.	Refuted
P160-DROPOUT035-FRESH-SEED-REPLICATION-PEMS08-20260923	The PEMS08 validation-MAE benefit of block_dropout_p=0.35 over 0.1 should replicate under three fresh seeds using the same frozen decision gate.	Gate-passed
P161-EXTERNAL-DEFAULT-AUDIT-20260923	At least one frozen regularization default in PGCRN or STAEformer should be revisable under the same validation-only three-seed protocol used for Orion.	Flat or inconclusive
P162-UNSCANNED-DEFAULTS-AUDIT-PEMS08-20260923	At least one of four previously untested frozen Orion defaults should produce a detectable validation-MAE change under a frozen one-factor audit.	Gate-passed
P165-REAL-AA-CALIBRATION-PEMS08-20260924	Under real seed-to-seed training noise, the frozen three-seed rule should trigger on same-configuration comparisons at a rate near the simulated 10.6% null rate (label band 0.056 to 0.156).	Gate-passed
P168-FRESH-SEED-REPLICATION-METRLA-GLR-AND-PEMSBAY-GRAPHON-20260924	The two original gate-pass results (METR-LA graph learning rate $\times 0.10$; PEMS-BAY graph on) should again pass the frozen three-seed rule when both arms are rerun on fresh seeds 3410-3412.	Gate-passed

Arm-level and audit crosswalk. The following entries split multi-arm contracts and name the audit records used for the numerical results. Their terminal labels are copied from the corresponding acceptance records (Table 5).

Table 5: Arm-level rulings and audit records for multi-arm contracts and diagnostics; outcomes follow accepted adjudications.

Registry ID	Description	Frozen label or status
–	METR-LA three-layer latent-separability diagnostic	INCONCLUSIVE; No demonstrated latent-separability gain
P126-EARLY	METR-LA early graph-activation arm	Earlier activation below gate
P126-GLR	METR-LA higher graph-learning-rate arm	Higher graph rate passes gate
P124-W96	PEMS08 width 96 arm	Gate-passed
P124-W128	PEMS08 width 128 arm	Gate-passed

Registry ID	Description	Frozen label or status (continued)
P141-p0.35	PEMS08 block dropout strength 0.35 arm	Stronger dropout gains on PEMS08
P145-later ruling	PEMS-BAY block dropout strength 0.35 arm	Stronger dropout flat or worse on PEMS-BAY
P147-review correction	METR-LA block dropout strength 0.35 arm; initial adjudicator read an unavailable field	Stronger dropout flat; initial earlier adjudication invalid due to a missing field
P135-METRLA replay	Post-selection METR-LA test receipt for validation-locked checkpoints	COMPLETE; no test-based selection or retry
P152B	Registry and contract-hash forensics	ACCEPTED audit
P154	Three-seed gate operating-characteristic and null-exceedance replay	ACCEPTED audit

D.1 CONTRACT-TO-ARM DISPLAY CROSSWALK

This dated registry contains 48 contracts. Of these, 27 contribute 44 of Figure 5's 48 displayed arm slots, while 21 have no displayed arm. The other 4 slots are diagnostic or arm-level records outside this registry snapshot. The first table gives each contract's hypothesis. Table 6 accounts for the displayed slots without treating an arm as a new contract.

Table 6: **Dated registry-to-figure mapping.** A dash means this contract has no Figure 5 slot. Additional slot sources outside the snapshot are described after the registry rows; their exact arm keys remain in the anonymous supplementary data.

Registry ID	Arm slot(s)
O4-NODE-CVAR-20260918	—
P097-V3-GRADGEOM-ATTRIBUTION-20260918	—
P098-GAIN-DECOMPOSITION-20260918	—
P099-ARMS-DIRECT-20260918	—
P100-ARMS-DIRECT-HARDENED-20260918	—
P101-SCDL-QUALGATE-20260918	—
P102-NSNM-20260918	—
P103-DIRPRIOR-20260918	—
P104-DIRPRIOR-MIN-20260918	—
P105-GAPLOC-20260918	—
P106-GAPLOC-BOUND-20260918	—
P107-INSTNORM-20260918	—
P108-INSTNORM-BOUND-20260918	—
P109-INSTNORM-CONTSUPP-20260918	—
P110-INSTNORM-RNGFROZEN-20260918	—
P112-IDENTITY-SHUFFLE-CONTROL-20260919	PEMS08: Node-identity shuffle
P113-GRAPH-CONTRIBUTION-20260919	PEMS08: Graph activation
P114-STOCHASTICITY-ISOLATION-20260919	PEMS08: Deterministic identity shrinkage
P115-METRLA-DROPOUT-CONTROL-20260919	METR-LA: Identity block dropout
P117-GRAPH-ON-METRLA-20260919	METR-LA: Graph activation

Registry ID	Arm slot(s) (continued)
P118-GRAPH-ON-PEMSBAY-20260920	PEMS-BAY: Graph activation
P120-DROPOUT-STRENGTH-METRLA-20260920	—
P121-BUDGET-MATCHED-SCHEDULE-PEMS08-20260920	PEMS08: Compressed training schedule
P124-CAPACITY-SCALING-PEMS08-20260920	PEMS08: Width 64 to 128
P125-DISABLED-SWITCHES-PEMS08-20260921	PEMS08: Adaptive spatial embeddings; PEMS08: Two-hop graph reach
P126-GRAPH-TRAINING-BUDGET-METRLA-20260921	METR-LA: Earlier graph activation; METR-LA: Higher graph learning rate
P129-BEST-ORION-METRLA-20260921	METR-LA: Width 64 to 128
P129C-CAPACITY-W128-PEMSBAY-20260922	PEMS-BAY: Width 64 to 128 with graph
P131-PNORM-LOSS-PEMS08-20260922	PEMS08: p-norm loss 1.25; PEMS08: p-norm loss 1.50
P132-CAPACITY-W192-PEMS08-20260922	PEMS08: Width 128 to 192
P133-GRAPH-LR-RETEST-PEMS08-20260922	PEMS08: Graph learning rate 0.10
P134-CALENDAR-BLOCK-DROPOUT-PEMS08-20260922	—
P136-BLOCK-DROPOUT-GRANULARITY-PEMS08-20260922	PEMS08: Per-node dropout mask
P137-GRAPH-LR-PEMSBAY-20260922	PEMS-BAY: Higher graph learning rate
P138-GRAPH-LR-UPPER-BOUND-METRLA-20260922	—
P140-QUANTILE-HEAD-PEMS08-20260923	PEMS08: Five-quantile output head
P141-DROPOUT-STRENGTH-PEMS08-20260923	PEMS08: Block dropout 0.20; PEMS08: Block dropout 0.35
P142-DROPOUT-STRENGTH-PEMSBAY-20260923	—
P145-DROPOUT-STRENGTH-PEMSBAY-20260923	PEMS-BAY: Block dropout 0.35
P147-DROPOUT-STRENGTH-METRLA-20260923	METR-LA: Block dropout 0.35
P148-DROPOUT-SWEEP-COMPLETION-PEMS08-20260923	PEMS08: Block dropout 0.50; PEMS08: Block dropout 0.65
P149-UNDERREG-AND-LR-PEMS08-20260923	PEMS08: Weight decay 1e-5; PEMS08: Weight decay 1e-4; PEMS08: Weight decay 1e-3; PEMS08: Main rate 0.003; PEMS08: Main rate 0.007; PEMS08: Main rate 0.010
P156-P157-GATE-CALIBRATION-PEMS08-20260923	—
P160-DROPOUT035-FRESH-SEED-REPLICATION-PEMS08-20260923	PEMS08: Block dropout 0.35, fresh seeds
P161-EXTERNAL-DEFAULT-AUDIT-20260923	PEMS08: PGCRN Weight decay 1e-3; PEMS08: PGCRN Weight decay 1e-2; PEMS08: STAEformer Dropout 0.20; PEMS08: STAEformer Dropout 0.35
P162-UNSCANNED-DEFAULTS-AUDIT-PEMS08-20260923	PEMS08: Batch size 32; PEMS08: Node embedding 32; PEMS08: Calendar embedding 8; PEMS08: Head depth 1
P165-REAL-AA-CALIBRATION-PEMS08-20260924	—

Registry ID	Arm slot(s) (continued)
P168-FRESH-SEED-REPLICATION-METRLA-GLR-AND-PEMSBAY-GRAPHON-20260924	METR-LA: Fresh graph rate 0.10; PEMS-BAY: Fresh graph on
Arm slots outside this dated registry snapshot	
—	METR-LA: Latent-separability probe
—	PEMS08: Adaptive instance norm; PEMS08: Static instance norm
—	PEMS08: Identity block dropout

Later cases outside Figure 5 and the 48-hypothesis registry. P169 and P170 are later protocol cases outside both the registry and the 48 display slots covered by the crosswalk above. P169 reports two fresh-seed known-harm reversals, and P170 reports six validation-only STID implementation-component verdicts. Appendix M gives their outcomes.

E Protocol and Orion adjudication details

E.1 Contract and protocol audit details

For the masked-loss comparison, the accepted record binds the frozen-contract digest `6784c7135082d651f41dd5ca751e57d0a450f91e032f42aa68971539284902b0` to the three-seed rule and decision table used for the modified-loss arms. The resulting evidence bundle contains the contract, run paths, per-seed metrics, discipline checks, adjudicated label, and any applicable exposure receipt. The digest authenticates the contract bytes. Separate records establish the chronology and scope of a prospective freeze.

The $p = 1$ masked-loss replay reproduced validation MAE values `13.405660828854629`, `13.367085227396428`, and `13.314486114179191` for seeds `3407`, `3408`, and `3409`. Its training curves and band-level checks also agreed bitwise. For the dropout-strength study, each treatment instead retained the identical initial-model-state SHA-256 and parameter count of its control because dropout is applied after initialization.

For the quantile-head comparison, the accepted validator read validation RMSE from the matched masked-L_1 control and returned $\Delta = -0.1555518204277142$, $s_{\max} = 0.1324826498778654$, and "no RMSE gain." This negative label was produced by the same frozen rule that could have accepted a positive result.

The final-recipe rule has SHA-256 `303d84ec0f4d782700e52663214cf635eabb4ff45b21f4c7d6ce3b8f64001e86`. The earlier METR-LA exposure receipt records three validation-locked checkpoints, one logical test opening, one evaluation pass per checkpoint, unchanged checkpoint hashes, and zero test-based selection or retry events. Exploratory graph-learning-rate results cannot enter the final recipe. The rule also requires disclosure of validation-comparison counts and dataset-specific recipe differences. If several PEMS08 components pass, it permits one prespecified combination arm but no subset search. The final test replay must be reported even when it is worse.

For the forecasting table, comparator admission requires a complete reported cell set, and the denominator remains fixed when a cell is unavailable. In the closed set without MetaDG, each of the 12 Orion-leading cells had a better excluded opponent. After restoring it, the count is 0 of 36 correlated dataset–horizon–metric display cells. This two-snapshot audit does not estimate the prevalence of incomplete comparisons in the literature.

The final dated registry contains 48 hypotheses, 9 refuted outcomes, 1 withdrawn before execution, and 15 gate-passed outcomes. Hash forensics found one truncated hash record, two contracts amended before any result, and 13 older contracts that lacked sidecars and were untracked at their original freeze. Of the final registry, 8 contracts had recorded pre-result submissions to public OpenTimestamps calendars: the later dropout sweep, the weight-decay/main-rate audit, and six subsequent contracts with individual calendar sidecars. The other 40 lack a pre-result external anchor. Some entered the later manifest, but that does not establish their earlier ordering. The two pre-result amendments, truncated hash, and 13 no-Git cases overlap those groups rather than adding to the

denominator. The calendar attestations had no recorded blockchain confirmation at the independent audit. Eight of the 48 contracts carry Bitcoin anchors whose block times precede their first results or no-run bounds; five of them do so by more than the two-hour tolerance (Appendix M.1). The contract-by-contract classification and source hashes are retained in the supplementary audit.

The final stoppage ruling classifies the 11 stoppages before a confirmatory conclusion as eight contract defects, one failed eligibility gate, and two non-evasion guards. Seven contract-defect cases have an explicit zero validation/test/GPU record. None of the 11 opened the test split. The failed eligibility gate and one non-evasion guard consumed GPU time, so the entire group is not described as pre-consumption. The individual cases are identified in the anonymous supplementary registry.

The Figure 5 plotting list was fixed at 05:17:01 on 2026-09-24. Hash and timestamp sidecars followed within seconds. It includes every listed treatment arm regardless of direction except the PEMS08 width-96 arm, one of the 12 numerical gate crossings, which is reported in Tables 20 and 27. Before that freeze, one node-embedding audit value and eight of nine weight-decay run values had already been viewed. The ninth run had completed but had not been read. Fresh-seed replication, external-model audit, and exploratory main-rate arms had no result then. The list's previously reported 05:30 timestamp and its blanket "not looked" descriptions were corrected by a separate erratum without changing the row set. The final registry contains 48 hypotheses. Its crosswalk maps 44 of the 48 displayed arm slots. 4 diagnostic slots lie outside this registry.

E.2 Exact gate-behavior replay values

The three-seed equal-variance normal simulation gave a null-trigger frequency of 0.10592. Its closed-form counterpart was 0.106462813. The empirical-residual replay gave 0.09502. Under the normal model, detection frequencies were 0.43928, 0.82334, and 0.97599 for effects of one, two, and three standard deviations. For 25 independent null comparisons, the simulated familywise frequency was 0.93913 under the normal model and 0.91759 under the empirical-residual replay. Shared-control comparisons do not satisfy the independence assumption. This multiplicity view was post hoc and did not change frozen labels. The real composite-degradation experiment's zero-impact control replay was bit-exact at validation MAE 13.405660828854629. Its ten unordered A/A partitions share the same six seeds, and one crossed the rule. They are dependent displays, not an empirical error-rate sample.

The later real-noise calibration used 18 validation seeds of one PEMS08 configuration. Among six fixed, disjoint three-seed blocks, 6 of 30 ordered control–arm comparisons crossed the frozen gate (20%). Across all 371,280 ordered disjoint three-versus-three splits, 38,603 crossed (10.397%), near the normal-model 10.592% reference. Under the null, the prespecified treatment mean was lower in 50% of these comparisons. Selecting the lower mean would name a winner in every untied comparison. The six-block view is higher because three blocks had very small within-block standard deviations (0.0045–0.0063). Both views reuse seeds, so neither yields an independent binomial interval or establishes a cross-dataset error rate. The earlier single A/A crossing is consistent with this observed null behavior.

E.3 Gate calibration sensitivity and family replay

The following checks are descriptive replays of the frozen decision rule. None changes an intervention label or estimates a universal error rate. Table 8 reports the semi-synthetic control, and Table 7 reports generator sensitivity.

Table 7: Sensitivity of the three-seed operational rule under stated paired normal or unit-variance Student-t generators. Variance ratio means $\mathrm{Var}(\mathrm{arm})/\mathrm{Var}(\mathrm{control})$; ρ is same-seed correlation. Each cell uses 200,000 simulated comparisons. A one-SD shift is defined relative to the control SD.

ρ	Variance ratio	t df	Null gate passes	One-SD detection
0.0	1	3	10.133%	59.191%
0.0	1	5	10.367%	49.941%
0.0	1	∞	10.600%	44.055%
0.0	2	3	10.171%	49.626%
0.0	2	5	10.418%	41.206%

ρ	Variance ratio	t df	Null gate passes	One-SD detection (continued)
0.0	2	∞	10.673%	36.220%
0.0	4	3	10.203%	39.296%
0.0	4	5	10.546%	32.762%
0.0	4	∞	11.029%	29.214%
0.3	1	3	7.355%	60.397%
0.3	1	5	7.491%	50.397%
0.3	1	∞	7.819%	43.889%
0.3	2	3	7.545%	49.803%
0.3	2	5	7.808%	40.634%
0.3	2	∞	8.112%	35.151%
0.3	4	3	8.127%	38.593%
0.3	4	5	8.392%	31.340%
0.3	4	∞	8.844%	27.601%
0.6	1	3	4.494%	63.111%
0.6	1	5	4.498%	52.405%
0.6	1	∞	4.763%	45.287%
0.6	2	3	4.912%	51.469%
0.6	2	5	5.217%	41.137%
0.6	2	∞	5.389%	34.859%
0.6	4	3	5.881%	38.371%
0.6	4	5	6.295%	30.478%
0.6	4	∞	6.798%	26.626%
0.9	1	3	1.231%	70.485%
0.9	1	5	1.263%	59.588%
0.9	1	∞	1.294%	52.629%
0.9	2	3	1.928%	55.899%
0.9	2	5	2.129%	44.805%
0.9	2	∞	2.314%	37.605%
0.9	4	3	3.672%	39.007%
0.9	4	5	3.981%	30.317%
0.9	4	∞	4.426%	25.529%

Table 8: **Semi-synthetic positive control.** A known additive shift is applied to one arm of disjoint three-seed comparisons drawn from 18 real PEMS08 validation seeds. Here s_{18} is their sample SD. Zero shift exactly reproduces the observed null split count. Detection here is not the power of any real intervention.

Known shift / s_{18}	Shift (MAE)	Gate passes	Semi-synthetic detection	Normal reference
0.00	0.00000000	38,603/371,280	10.397%	10.592%
0.25	0.00815886	61,790/371,280	16.642%	—
0.50	0.01631772	90,742/371,280	24.440%	—
1.00	0.03263544	151,466/371,280	40.796%	43.928%
2.00	0.06527088	301,664/371,280	81.250%	82.334%
3.00	0.09790632	368,480/371,280	99.246%	97.599%

Positive same-seed correlation lowers the null trigger frequency in this generator. The table does not estimate empirical cross-arm correlation.

Table 9 reports the shared-control family replay.

E.4 Decision-rule comparison and controlled negative case

The separate rule comparison uses the same three-seed, equal-variance normal generator as the operating-characteristic replay. Declaring the smaller mean the leader has a much higher null-trigger rate than the contracted rule. The one-sided Welch reference uses nominal $\alpha = 0.05$; its simulated null-trigger rate is 3.968% with three seeds per arm. The rules have different zero-effect trigger rates, so their detection curves do not compare efficiency at matched error rates. These are properties of specified decision rules under specified draws, not tests of published papers or revisions to earlier labels.

Table 9: **Descriptive family-level global-null replay.** Each physical run group receives one independent normal noise triple per simulation, reused wherever that group enters another decision. Cross-metric reuse is an explicit perfect-dependence sensitivity, not an estimated covariance. These probabilities do not re-adjudicate frozen labels or establish familywise error control.

Family-level quantity	Accepted replay value
Eligible arm-level decisions	45
Observed gate passes	12
Unique physical run groups	53
Shared-control groups	6
Final-registry rows excluded by rule or scope	5
Mean gate passes under the global null	4.80
Probability of at least one gate pass	96.023%
Probability of at least the observed count	0.065

Table 10 reports the controlled case. The SEIR experiment used a frozen CPU/FP32 protocol with

Table 10: **Controlled SEIR falsification under its CPU/FP32 protocol.** For each seed, Δ is Orion's structure area under the receiver operating characteristic curve (AUROC) minus the best of the random, fully connected, and correlation-only controls; the table gives its across-seed mean and deterministic 5,000-resample 95% bootstrap interval. Q2 requires Spearman ≥ 0.50 and normalized MAE ≤ 0.25. Q3 requires I-NRMSE ≤ 0.50, effect-I NRMSE ≤ 0.50, and no more than a 5% I-NRMSE disadvantage to the best control. Q2 and Q3 count passing seeds. The medium interval is wholly negative; the easy and hard intervals span zero. This case is separate from gate calibration. The nominal 95% bootstrap intervals summarize only three seeds per tier and are descriptive; the medium-tier interval's exclusion of zero is not treated as calibrated evidence.

Tier	Mean Δ	95% interval	Q2	Q3
Easy	-0.0497	[-0.2655, 0.0610]	0/3	0/3
Medium	-0.0776	[-0.1731, -0.0094]	0/3	0/3
Hard	-0.0193	[-0.0727, 0.0145]	0/3	0/3

a 0.60/0.20/0.20 split, 12-step input/output windows, and three seeds per tier. Easy and medium observe every region; hard leaves 25% of regions unobserved, with Boolean input/target validity masks. Scaling uses factual training windows only; contact truth, clean states, and interventions remain hidden until post-fit evaluation. The negative verdict holds at all tiers, while only the medium interval remains wholly below zero. No positive recovery control was included, so this case establishes that the protocol can retain an adverse result, not that it can recover known structure when present.

E.5 Cross-dataset dropout-strength boundary

The accepted PEMS-BAY dropout-strength treatment at 0.35 worsened validation MAE relative to its 0.1 control and received the frozen "already near optimal or worse" decision. The METR-LA treatment at 0.35 was flat after independent review corrected an adjudicator field-mapping defect: the original script asked for a graph-activity field emitted only by the PEMS-BAY runner and therefore returned an inconclusive discipline-failure status. Under the pre-existing repair-on-unavailable-field rule, the review mapped that check to the recorded activation epoch (31), graph-rate ratio (0.1), and positive graph-training epochs before selection (23/36/23), then accepted the flat dropout-strength decision. The unchanged adjudicator still emits its original inconclusive status; the reviewed flat ruling is reported separately, and neither the frozen contract nor adjudicator was changed. The PEMS08 gain was replicated with fresh seeds: control mean 13.423255392876792, treatment mean 13.292825205103789, difference 0.13043018777300297 MAE. The separate A/A block shift was 0.060844669400042406 MAE. The six-seed pooled difference 0.10903198237713063 MAE is descriptive and changes no frozen decision. The cross-dataset outcomes limit the replicated effect to the tested PEMS08 recipe. The formal three-arm weight-decay audit and exploratory three-arm main-rate audit were both flat under their respective rules. The four previously unscanned defaults were audited under the earlier 0.10 block-dropout base, not the final 0.35 recipe. The frozen control mean was 13.362411 validation MAE. Table 11 reports the four audited defaults. Halving batch size

Table 11: **Four frozen default audits on PEMS08 under the earlier 0.10 block-dropout base.** A positive gain means a lower validation MAE after the single change. Null exceedance probabilities are descriptive.

Changed default	Frozen decision	Arm MAE	Gain MAE	Null exceedance
Batch size 32	Worse	13.424088	-0.061677	0.0603
Node embedding 32	Flat	13.350715	0.011695	0.3612
Calendar embedding 8	Flat	13.379038	-0.016627	0.3085
Head depth 1	Flat	13.385321	-0.022910	0.2488

crossed the frozen rule in the worse direction by $1.33 s_{\max}$, comparable to the observed A/A block shift of 0.061 MAE. This is weak evidence that the 64 default is not harmful. It does not establish optimality or a general batch-size effect. None of these audits revised the final recipe or opened test.

E.6 Capacity, graph-rate, and dropout response curves

Figure 6 shows the capacity and graph-rate responses, Figure 7 shows dropout responses, and Figure 8 shows the gate checks.

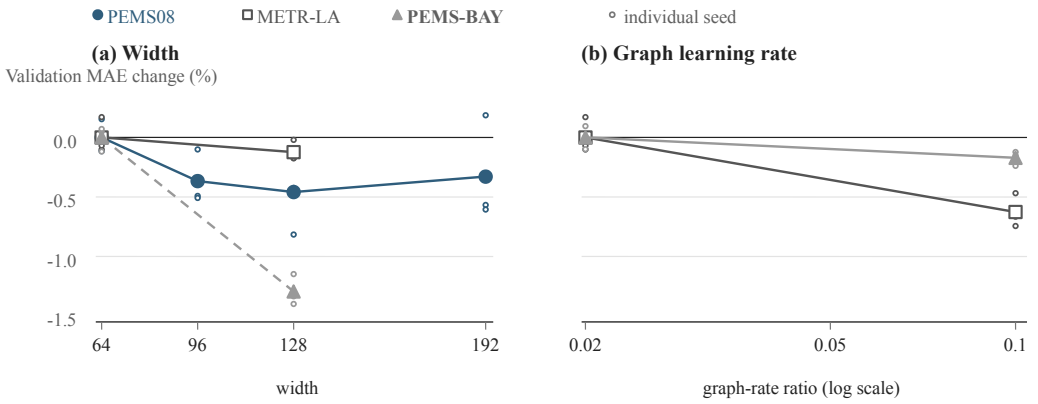

Figure 6: Accepted validation MAE responses, normalized within each panel to each dataset's first plotted reference. Small rings represent individual seeds. Larger dataset-shaped marks are means. In (a), the dashed PEMS-BAY connector joins a graph-off width-64 control to a graph-on width-128 arm, so the change cannot isolate width. The width-192 frozen decision compared against width 128, although the display uses width 64 as its reference. In (b), the graph-rate axis is logarithmic. The grid also contained ratios 0.3 (one completed seed) and 1.0 (none). Neither has a three-seed verdict or a plotted mean.

F Recorded historical test-exposure ledger

All times are given in UTC; rows follow the supplementary ledger order. The accepted ledger records 71 logical events. These comprise 67 raw test evaluations, 1 attempt aborted before a forward pass, and 3 test-result exposures. Of these events, 56 have documented times and 15 have no verified date. Exact event IDs and evidence pointers remain in the anonymous supplementary ledger, whose row order is identical. Each event groups a model/recipe, dataset, checkpoint cohort, and logical pass. Checkpoint/run test passes total at least 241. Later-use flags are 15 YES, 6 NO, and 50 UNKNOWN. YES denotes documented later use, including hypothesis choice. It does not imply that the final recipe was selected on test. Dataset, recipe, timing, purpose, later-use flag, and raw-pass bounds retain the source ledger meanings. Table 12 lists all 71 recorded events.

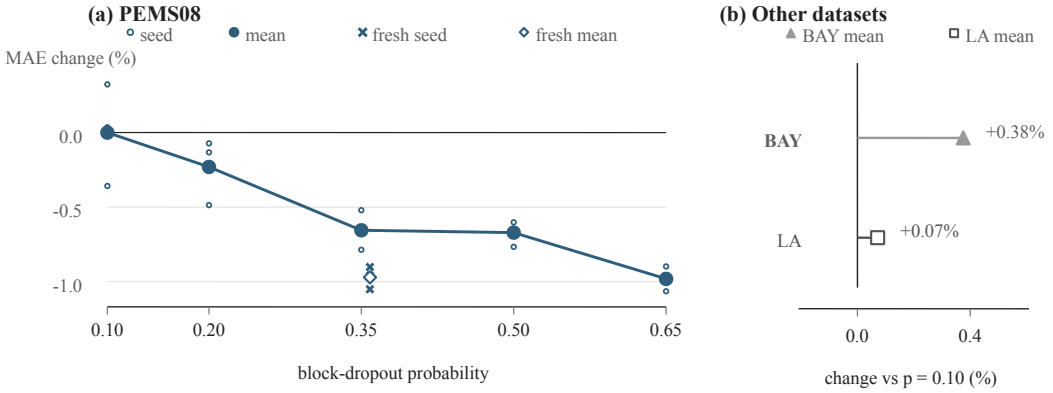

Figure 7: **Within-dataset block-dropout responses.** PEMS08 held the graph off. METR-LA and PEMS-BAY used their distinct graph-on recipes. Original seed rings and means are relative to each dataset's 0.10 control. The fresh PEMS08 crosses and diamond at 0.35 use their own fresh 0.10 control. Values at 0.50 and 0.65 were judged against 0.35, although the curve is displayed relative to 0.10. Means declined through 0.65, but the extra gain did not pass the frozen rule ($0.953 s_{\max}$ at 0.65). PEMS-BAY worsened at 0.35. METR-LA was flat.

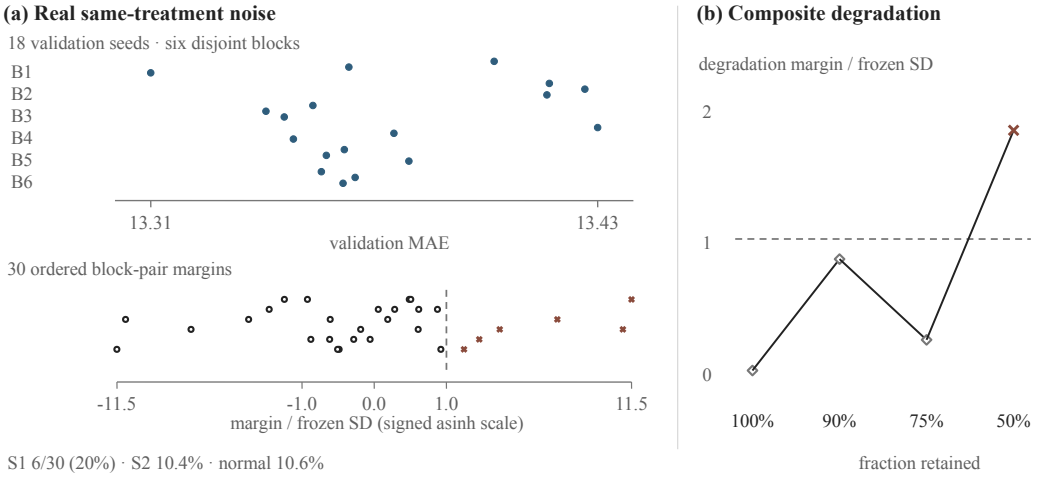

Figure 8: **Gate-check detail.** (a) All 18 real same-treatment validation seeds in six fixed blocks and all 30 ordered block-pair margins, shown on a signed inverse-hyperbolic-sine axis so the frozen one-$s_{\max}$ threshold remains visible beside large margins. Open circles mark nontriggers and terra-cotta crosses mark triggers. Six block pairs cross (20%). Exhaustive dependent three-versus-three partitions with disjoint arms cross in 38,603/371,280 (10.4%), near the simulated 10.6% reference. Reused seeds preclude an independent error-rate interval. (b) Composite degradation with 90%, 75%, and 50% of training data retained. The response is nonmonotone and only 50% crosses. Sample count and optimizer-update count change together.

Table 12: Historical test-exposure ledger of 71 logical events; each row records timing, purpose, later use, and the raw-pass bound.

Event	Exposure record
Exposure 01	**kind**: Raw test evaluation; **time**: 2026-09-05T20:31:54Z; **time basis**: task-registry update time (upper bound); **dataset**: PEMS04; **recipe**: final architecture; exact recipe UNKNOWN; 3 seeds; **trigger**: RUN-FINAL-PEMS04-x3; **purpose**: other; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 1; **note**: the task-registry decision field records that the test result was used to judge the final architecture gate.
Exposure 02	**kind**: Test-result exposure; **time**: 2026-09-13T10:53:00Z; **time basis**: task-registry update time (upper bound); **dataset**: PEMS08; **recipe**: historical multi-recipe summaries; **trigger**: historical accuracy documentation review; **purpose**: other; **later use**: NO; **raw passes lower bound**: 0; **raw file opens**: 0; **evidence**: supplementary exposure ledger, row 2; **note**: Startup documentation exposure only; task failed immediately with no validation/test split read.
Exposure 03	**kind**: Raw test evaluation; **time**: 2026-09-19T00:00:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: STAEformer; 3 seeds; **trigger**: B2A-STAEFORMER-METRLA-20260918; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 3; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 04	**kind**: Raw test evaluation; **time**: 2026-09-19T02:50:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: Orion recipe accepted on 19 September 2026; 3 checkpoints; **trigger**: O3-ORION-PEMS08-20260919; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 4; **note**: Task-local test selection/retry=0 does not prove absence of later global use.
Exposure 05	**kind**: Raw test evaluation; **time**: 2026-09-19T07:05:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: STAEformer; 3 seeds; **trigger**: B2B-STAEFORMER-PEMSBAY-20260919; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 5; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 06	**kind**: Raw test evaluation; **time**: 2026-09-19T07:40:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: PGCRN; 3 seeds; **trigger**: B1-PEMS08-BASELINE-20260918; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 6; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 07	**kind**: Raw test evaluation; **time**: 2026-09-19T07:40:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: ST-SSDL; 3 seeds; **trigger**: B1-PEMS08-BASELINE-20260918; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 7; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 08	**kind**: Raw test evaluation; **time**: 2026-09-19T08:30:00Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: Orion recipe accepted on 19 September 2026; 3 checkpoints; **trigger**: O3-ORION-METRLA-20260919; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 8; **note**: Task-local test selection/retry=0 does not prove absence of later global use.
Exposure 09	**kind**: Raw test evaluation; **time**: 2026-09-19T10:59:18Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: STAEformer; 3 seeds; **trigger**: B2C-STAEFORMER-PEMS08-20260919; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 9; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 10	**kind**: Raw test evaluation; **time**: 2026-09-19T17:19:57Z; **time basis**: completion upper bound recorded in acceptance; **dataset**: METR-LA; **recipe**: AGCRN; 3 seeds; **trigger**: B3-AGCRN-THREE-DATASETS-20260919; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 10; **note**: Protocol proves validation-only selection and one test report; later global use remains UNKNOWN.

Event	Exposure record (continued)
Exposure 11	**kind**: Raw test evaluation; **time**: 2026-09-19T17:19:57Z; **time basis**: completion upper bound recorded in acceptance; **dataset**: PEMS08; **recipe**: AGCRN; 3 seeds; **trigger**: B3-AGCRN-THREE-DATASETS-20260919; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 11; **note**: Protocol proves validation-only selection and one test report; later global use remains UNKNOWN.
Exposure 12	**kind**: Raw test evaluation; **time**: 2026-09-19T17:19:57Z; **time basis**: completion upper bound recorded in acceptance; **dataset**: PEMS-BAY; **recipe**: AGCRN; 3 seeds; **trigger**: B3-AGCRN-THREE-DATASETS-20260919; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 12; **note**: Protocol proves validation-only selection and one test report; later global use remains UNKNOWN.
Exposure 13	**kind**: Raw test evaluation; **time**: 2026-09-19T23:03:32Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: MetaDG; 3 seeds; **trigger**: B1-PEMS08-BASELINE-20260918; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 13; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 14	**kind**: Raw test evaluation; **time**: 2026-09-20T07:52:28Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: AutoSTF-NoNAS†; 3 seeds; **trigger**: B1-AUTOSTF-PEMS08-FIX-20260920; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 14; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 15	**kind**: Raw test evaluation; **time**: 2026-09-20T22:06:10Z; **time basis**: recorded time; **dataset**: METR-LA; **recipe**: TESTAM; 3 seeds; **trigger**: B4 TESTAM expansion; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 15; **note**: The test head-to-head is explicitly recorded as a reason for P129.
Exposure 16	**kind**: Raw test evaluation; **time**: 2026-09-21T03:43:02Z; **time basis**: recorded time; **dataset**: METR-LA; **recipe**: Orion; existing 3-checkpoint cohort; **trigger**: per-horizon diagnostic evaluation, second version; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 16; **note**: Test head-to-head explicitly motivated P129.
Exposure 17	**kind**: Raw test evaluation; **time**: 2026-09-21T03:43:02Z; **time basis**: recorded time; **dataset**: PEMS08; **recipe**: Orion; existing 3-checkpoint cohort; **trigger**: per-horizon diagnostic evaluation, second version; **purpose**: diagnosis; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 17
Exposure 18	**kind**: Raw test evaluation; **time**: 2026-09-21T03:43:02Z; **time basis**: recorded time; **dataset**: PEMS08; **recipe**: STAEformer; existing 3-checkpoint cohort; **trigger**: per-horizon diagnostic evaluation, second version; **purpose**: diagnosis; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 18
Exposure 19	**kind**: Raw test evaluation; **time**: 2026-09-21T03:43:02Z; **time basis**: recorded time; **dataset**: PEMS-BAY; **recipe**: STAEformer; existing 3-checkpoint cohort; **trigger**: per-horizon diagnostic evaluation, second version; **purpose**: diagnosis; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 19
Exposure 20	**kind**: Raw test evaluation; **time**: 2026-09-21T13:10:51Z; **time basis**: recorded time; **dataset**: PEMS-BAY; **recipe**: TESTAM; 3 seeds; **trigger**: B4 TESTAM expansion; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 20; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 21	**kind**: Raw test evaluation; **time**: 2026-09-21T15:39:01.868091Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: Orion W128_GRAPH_OFF; P124 checkpoints; **trigger**: P130-BEST-ORION-PEMS08-20260921; **purpose**: post-selection; **later use**: NO; **raw passes lower bound**: 3; **raw file opens**: 8; **evidence**: supplementary exposure ledger, row 21; **note**: Source states the table replacement would occur even if test were worse.

Event	Exposure record (continued)
Exposure 22	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: MetaDG; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 22; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 23	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: MetaDG; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 23; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 24	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: Orion; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 24; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 25	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: PGCRN; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 25; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 26	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: PGCRN; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 26; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 27	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: PGCRN; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 27; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 28	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: STAEformer; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 28; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 29	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: ST-SSDL; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 29; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 30	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: ST-SSDL; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 30; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.

Event	Exposure record (continued)
Exposure 31	**kind**: Raw test evaluation; **time**: 2026-09-21T15:43:47.259291Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: ST-SSDL; 3 checkpoints; **trigger**: per-horizon diagnostic evaluation, third version, 21 September 2026; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 31; **note**: YES is campaign-level: the ten-cell per-horizon test campaign informed later weakness localization/calibration design; marginal attribution to one cell is unavailable.
Exposure 32	**kind**: Raw test evaluation; **time**: 2026-09-21T21:31:17.486503Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: TESTAM-PEMS08-Transfer†@20ee7815; 3 seeds; **trigger**: B6-TESTAM-PEMS08-20260921; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 32; **note**: Task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 33	**kind**: Raw test evaluation; **time**: 2026-09-22T12:24:04.419293Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: AutoSTF-NoNAS†; supplementary post-selection replay; 3 seeds; **trigger**: baseline supplementary replay: AutoSTF-NoNAS†—PEMS08; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 33; **note**: Explicitly distinct from the earlier P088/B1 reporting pass; later global use remains UNKNOWN.
Exposure 34	**kind**: Raw test evaluation; **time**: 2026-09-22T12:24:06.928911Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: AutoSTF-NoNAS†; supplementary post-selection replay; 3 seeds; **trigger**: baseline supplementary replay: AutoSTF-NoNAS†—METR-LA; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 34; **note**: Explicitly distinct from the earlier P088/B1 reporting pass; later global use remains UNKNOWN.
Exposure 35	**kind**: Raw test evaluation; **time**: 2026-09-22T12:24:11.365061Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: AutoSTF-NoNAS†; supplementary post-selection replay; 3 seeds; **trigger**: baseline supplementary replay: AutoSTF-NoNAS†—PEMS-BAY; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 35; **note**: Explicitly distinct from the earlier P088/B1 reporting pass; later global use remains UNKNOWN.
Exposure 36	**kind**: Raw test evaluation; **time**: 2026-09-22T18:13:22.366766Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: STID; 3 seeds; **trigger**: baseline formal acceptance: STID—PEMS08; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 36; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 37	**kind**: Raw test evaluation; **time**: 2026-09-22T18:13:22.373963Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: STID; 3 seeds; **trigger**: baseline formal acceptance: STID—METR-LA; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 37; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 38	**kind**: Raw test evaluation; **time**: 2026-09-22T18:13:22.378197Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: STID; 3 seeds; **trigger**: baseline formal acceptance: STID—PEMS-BAY; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 38; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 39	**kind**: Raw test evaluation; **time**: 2026-09-22T20:19:40.773319Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: Graph-WaveNet; 3 seeds; **trigger**: baseline formal acceptance: Graph-WaveNet—PEMS08; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 39; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.

Event	Exposure record (continued)
Exposure 40	**kind**: Raw test evaluation; **time**: 2026-09-22T20:19:40.781386Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: Graph-WaveNet; 3 seeds; **trigger**: baseline formal acceptance: Graph-WaveNet—METR-LA; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 40; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 41	**kind**: Raw test evaluation; **time**: 2026-09-22T20:19:40.786705Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: Graph-WaveNet; 3 seeds; **trigger**: baseline formal acceptance: Graph-WaveNet—PEMS-BAY; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 41; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 42	**kind**: Test-result exposure; **time**: 2026-09-22T13:35:00Z; **time basis**: acceptance time; **dataset**: MULTI; **recipe**: three datasets / multiple models; **trigger**: main-table source switch and integrity gate; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 0; **raw file opens**: 0; **evidence**: supplementary exposure ledger, row 42
Exposure 43	**kind**: Raw test evaluation; **time**: 2026-09-23T00:56:20.773951Z; **time basis**: recorded completion time; **dataset**: METR-LA; **recipe**: Orion P126-GLR; 3 checkpoints; **trigger**: P135-ORION-BEST-PROPAGATION-20260922; **purpose**: post-selection; **later use**: NO; **raw passes lower bound**: 3; **raw file opens**: 6; **evidence**: supplementary exposure ledger, row 43
Exposure 44	**kind**: Raw test evaluation; **time**: 2026-09-23T00:58:16.926840Z; **time basis**: recorded completion time; **dataset**: PEMS-BAY; **recipe**: Orion P137 W64 + graph_lr_ratio=0.1; 3 checkpoints; **trigger**: P135-ORION-BEST-PROPAGATION-20260922; **purpose**: post-selection; **later use**: NO; **raw passes lower bound**: 3; **raw file opens**: 6; **evidence**: supplementary exposure ledger, row 44
Exposure 45	**kind**: Raw test evaluation; **time**: 2026-09-23T04:18:07.222508Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: D2STGNN; 3 seeds; **trigger**: baseline formal acceptance: D2STGNN—METR-LA; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 45; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 46	**kind**: Test-result exposure; **time**: 2026-09-23T01:40:37Z; **time basis**: freeze timestamp; **dataset**: MULTI; **recipe**: METR-LA P126 and PEMS-BAY P137 aggregate summaries; **trigger**: P143 case-selection freeze; **purpose**: other; **later use**: UNKNOWN; **raw passes lower bound**: 0; **raw file opens**: 0; **evidence**: supplementary exposure ledger, row 46; **note**: Source proves no P143 rule selection from test; future global use cannot be excluded.
Exposure 47	**kind**: Raw test evaluation; **time**: 2026-09-23T10:44:41.210726Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: D2STGNN; 3 seeds; **trigger**: baseline formal acceptance: D2STGNN—PEMS08; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 47; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 48	**kind**: Raw test evaluation; **time**: 2026-09-23T14:31:25.369227Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: D2STGNN; 3 seeds; **trigger**: baseline formal acceptance: D2STGNN—PEMS-BAY; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 48; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 49	**kind**: Raw test evaluation; **time**: 2026-09-24T09:10:51.062515Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: HimNet; 3 seeds; **trigger**: baseline formal acceptance: HimNet—PEMS-BAY; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 49; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 50	**kind**: Raw test evaluation; **time**: 2026-09-24T09:10:51.066569Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: HimNet; 3 seeds; **trigger**: baseline formal acceptance: HimNet—METR-LA; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 50; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.

Event	Exposure record (continued)
Exposure 51	**kind**: Attempt aborted before forward pass; **time**: 2026-09-24T01:37:51Z; **time basis**: armed_at / acceptance exposure history; **dataset**: PEMS08; **recipe**: Orion final PEMS08 recipe P141/P_035; 3 checkpoints; **trigger**: P146B; **purpose**: post-selection; **later use**: NO; **raw passes lower bound**: 0; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 51; **note**: Attempt aborted before any test-batch forward pass; zero metrics and zero test values were produced. Recorded as an attempt, not RAW_TEST_EVAL.
Exposure 52	**kind**: Raw test evaluation; **time**: 2026-09-24T02:22:44Z; **time basis**: acceptance exposure history completion time; **dataset**: PEMS08; **recipe**: Orion final PEMS08 recipe P141/P_035; 3 checkpoints; **trigger**: P146C; **purpose**: post-selection; **later use**: NO; **raw passes lower bound**: 3; **raw file opens**: 8; **evidence**: supplementary exposure ledger, row 52; **note**: The second attempt was the only completed evaluation; result was used as-is and not for selection or retry.
Exposure 53	**kind**: Raw test evaluation; **time**: 2026-09-24T13:22:57.382437Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: HimNet; 3 seeds; **trigger**: baseline formal acceptance: HimNet—PEMS08; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 53; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 54	**kind**: Raw test evaluation; **time**: 2026-09-24T15:24:37.411737Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS08; **recipe**: TrendGCN; 3 seeds; **trigger**: baseline formal acceptance: TrendGCN—PEMS08; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 54; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 55	**kind**: Raw test evaluation; **time**: 2026-09-24T16:10:24.950904Z; **time basis**: acceptance verification time (upper bound); **dataset**: PEMS-BAY; **recipe**: TrendGCN; 3 seeds; **trigger**: baseline formal acceptance: TrendGCN—PEMS-BAY; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 55; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 56	**kind**: Raw test evaluation; **time**: 2026-09-24T17:14:43.584495Z; **time basis**: acceptance verification time (upper bound); **dataset**: METR-LA; **recipe**: TrendGCN; 3 seeds; **trigger**: baseline formal acceptance: TrendGCN—METR-LA; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 56; **note**: Acceptance proves task-local selection/retry=0; later global use remains UNKNOWN.
Exposure 57	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS-BAY; **recipe**: Orion recipe accepted on 19 September 2026; 3 checkpoints; **trigger**: O3-ORION-PEMSBAY-20260919; **purpose**: post-selection; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 57; **note**: Task-local test selection/retry=0 does not prove absence of later global use.
Exposure 58	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: METR-LA; **recipe**: AutoSTF-NoNAS†; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 58; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 59	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS-BAY; **recipe**: AutoSTF-NoNAS†; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 59; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 60	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: METR-LA; **recipe**: MetaDG; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 60; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 61	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS-BAY; **recipe**: MetaDG; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 61; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.

Event	Exposure record (continued)
Exposure 62	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: METR-LA; **recipe**: PGCRN; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 62; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 63	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS-BAY; **recipe**: PGCRN; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 63; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 64	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: METR-LA; **recipe**: ST-SSDL; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 64; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 65	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS-BAY; **recipe**: ST-SSDL; 3 seeds; **trigger**: P088 lineage; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 65; **note**: P088 proves task-local selection/retry=0; later global use is not fully recorded.
Exposure 66	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: AutoSTF-NoNAS†; 3 seeds; **trigger**: modern baseline; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 66
Exposure 67	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: MetaDG; 3 seeds; **trigger**: modern baseline; **purpose**: main-table reporting; **later use**: UNKNOWN; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 67
Exposure 68	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: Orion multi-recipe historical campaign (multiple variants with graph on/off); **trigger**: early accuracy line; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: $\geq$38; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 68; **note**: Campaign-level lower bound; source calls the conclusions test-informed/exploratory and records downstream direction choice.
Exposure 69	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: PGCRN; 3 seeds; **trigger**: modern-baseline diagnostic; **purpose**: diagnosis; **later use**: YES; **raw passes lower bound**: 3; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 69; **note**: PGCRN test MAE was recorded as an Orion accuracy target.
Exposure 70	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: STAEformer; n=1; **trigger**: modern-baseline diagnostic; **purpose**: diagnosis; **later use**: UNKNOWN; **raw passes lower bound**: 1; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 70
Exposure 71	**kind**: Raw test evaluation; **time**: UNKNOWN; **time basis**: UNKNOWN; **dataset**: PEMS08; **recipe**: ST-SSDL; 10-seed campaign; **trigger**: modern baseline; **purpose**: diagnosis; **later use**: UNKNOWN; **raw passes lower bound**: 10; **raw file opens**: UNKNOWN; **evidence**: supplementary exposure ledger, row 71; **note**: Earlier n=1 value is treated as a restatement/subset to avoid double counting.

G ORION ARCHITECTURE AND IMPLEMENTATION

The implementation has a fixed history length $T = 12$ and forecast horizon $H = 12$. It uses three registered graph lags $K = 3$, batch size B, and dataset-specific node count N. Its hidden width is $D = W \in \{64, 128\}$. The observed signal X and input-validity mask M have shape $[B, T, N, 1]$. Known input calendar features have shape $[B, T, 2]$. Origin time-of-day and day-of-week indices condition the encoder. Future calendar indices condition the forecast head. Node identifiers index learned embeddings and never mix observations across nodes.

Evaluation data and checkpoint selection. All models admitted to the forecasting ledger use the same stored split arrays within each dataset, 12 input steps, 12 output steps, and windows contained

entirely within one split. Table 13 gives the raw time-step boundaries. A single scalar z-score transform is fitted on the training array only; zeros participate in that fit under the BasicTS convention. Evaluation retains a target exactly when it is finite and nonzero, and applies the same Boolean mask to MAE, RMSE, and WAPE.

Table 13: **Traffic-data split boundaries shared by all admitted models.** Rows are raw time steps before 12-input, 12-output window construction.

Dataset	Train	Validation	Test	Ratio
PEMS08	10,713	3,571	3,572	60/20/20
METR-LA	23,990	3,427	6,855	70/10/20
PEMS-BAY	36,481	5,211	10,424	70/10/20

Data processing and checkpoint evaluation are shared, while optimization budgets remain model-specific. Orion runs the frozen 150-epoch schedule without early stopping and selects the raw or exponential-moving-average checkpoint with the lowest full-validation original-scale masked MAE under the frozen automatic rule. The fixed-recipe rerun harness caps AGCRN, Graph-WaveNet, and STAEformer at 100, 100, and 300 epochs with patience 50, 100, and 20, respectively. The expanded official-model runs map an initial 10-epoch probe to a frozen cap of 50, 100, or 200 epochs, retain each model's official patience, and select the lowest full-validation MAE checkpoint. Optimizer, batch size, learning-rate schedule, and patience therefore follow the frozen model recipe. Test arrays remain closed until checkpoint selection, after which each selected checkpoint receives one reporting pass.

Node-local encoder and training mask. The encoder concatenates the signal, validity indicator, and two calendar channels at each node and time, then projects four channels to 32 token channels and adds 12 learned positions. Four conditioned dilated token blocks use dilations $(1, 2, 4, 8)$. The condition vector has 64 node-identity, 32 time-of-day, and 32 day-of-week coordinates. During a training forward call, a single Bernoulli draw $Z \sim \mathrm{Bernoulli}(p)$ replaces the identity slice of this condition with zero when $Z = 1$. Evaluation does not draw or mask. A history projection and three residual fusion blocks produce node-local context $C \in \mathbb{R}^{B\times N\times D}$. The wrapped base encoder also supplies current and lagged states $H^{(0)} \in \mathbb{R}^{B\times N\times D}$ and $H^{(\ell)} \in \mathbb{R}^{B\times N\times K\times D}$. These are distinct outputs of the same temporal forward call, and the diagram combines them into one encoder box.

Directed influence and propagation. For each registered candidate edge $e = (s \to r)$ and sample b, the influence module produces a gate g_{be}, a direction weight d_{be}, and lag simplex $\gamma_{be\ell}$, where $\sum_{\ell=1}^{K} \gamma_{be\ell} = 1$. The effective propagation coefficient is $a_{be\ell} = g_{be} d_{be} \gamma_{be\ell}$. With projected source states $V(H^{(\ell)}_{bs\ell})$ and inverse-degree scale D_r^{-q}, where D_r is the registered in-degree of receiver r, the first-hop message is

$$M^{(1)}_{br\ell} = D_r^{-q} \sum_{e:s\to r} a_{be\ell}\, V(H^{(\ell)}_{bs\ell}), \qquad M^{(1)} \in \mathbb{R}^{B\times N\times K\times D}. \tag{2}$$

In the selected configurations, $q = 1$ is a fixed degree exponent. $g_{be} \in (0, 1)$ is a sigmoid edge gate and $d_{be} \in [0, 1]$ a direction weight. Both depend on the sample and edge and come from separately parameterized heads whose weights are fitted per graph-on dataset. The graph-on configurations add a second hop: a bias-free rank-16 down/up transform of $\mathrm{stopgrad}(M^{(1)})$ is routed over the same directed edges, then bounded per node and lag to have norm no larger than the detached first hop. Its result $M^{(2)}$ has the same shape. Horizon-specific coefficients $\rho_{h\ell} = K\,\mathrm{softmax}_\ell(R_{h\ell})$ produce $M_{bhr:} = \sum_\ell \rho_{h\ell}(M^{(1)}_{br\ell} + M^{(2)}_{br\ell})$, shape $[B, H, N, D]$. No separate second-order edge is learned or exported. With graph disabled, the output correction below is exactly zero, even though unused graph parameters remain instantiated.

Forecast readout and correction. The horizon-conditioned readout F uses learned horizon queries, future calendar embeddings, a residual hidden projection, and a scalar output projection. A separate direct projection of C gives a node-local output term. Writing S for the local propaga-

tion state, the selected model returns

$$Y_{\text{local}} = F(S) + \text{Direct}(C), \qquad \kappa_{rh} = \frac{\theta_{rh}}{1 + |\theta_{rh}|},$$
$$\widehat{Y} = Y_{\text{local}} + I_{\text{graph}}\, \kappa_{rh} \big[F_h(S + M) - F_h(S)\big]. \tag{3}$$

Here θ_{rh} is a learned receiver–horizon parameter, so $\kappa_{rh} \in (-1, 1)$ scales the graph correction for receiver r at horizon h; F_h denotes the horizon-specific evaluation of the same forecast-head weights and I_{graph} is the graph-on switch. The source implements the local and corrected evaluations inside its output-residual hook. The head has depth two in all three selected configurations. The formula describes predictive propagation weights and does not identify effects of physical interventions.

Figure 9 expands the implementation modules.

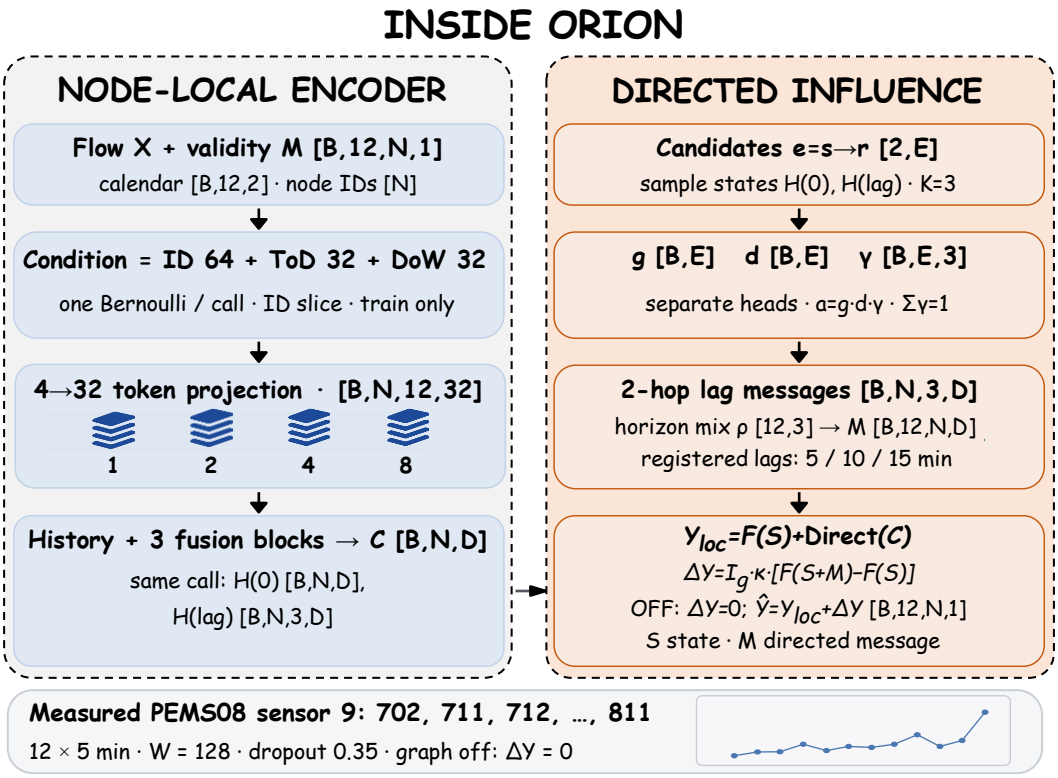

Figure 9: **Module detail behind Fig. 3.** The left panel expands the node-local temporal path and its training-only identity mask. ToD and DoW denote time of day and day of week. B, N, D, and E denote batch size, nodes, latent width, and candidate edges. The right panel expands directed influence and optional forecast correction. The three 5/10/15-minute labels are registered lag supports, not measured effects for the example sensor. The bottom trace is an observed PEMS08 sensor-9 history. Its selected graph-off configuration makes the graph correction exactly zero.

Parameter accounting. Table 14 gives counts recomputed as the sum of tensor element counts in one selected checkpoint for each dataset. The "Influence" and "Output residual" columns include instantiated parameters even in the graph-off PEMS08 model. Block dropout adds no parameter. All columns sum to the total in the corresponding accepted result.

Figure-to-code map. The PEMS08 dropout guard is the sole altered training-source file against its pinned width-128 parent. METR-LA and PEMS-BAY use their run-pinned width-64 trees. The graph learning-rate schedule is controlled by each dataset run's declared candidate and its options implementation. The accompanying source code gives exact file and symbol locators. Table 15 maps diagram outputs to implementation roles and is not an additional scientific claim.

Table 14: **Parameter counts of the selected Orion configurations.**

Dataset	Temporal	Influence	Propagation	Forecast head	Output residual	Total
PEMS08	321,520	231,045	32,768	105,217	1,742	692,292
METR-LA	295,152	58,181	8,192	36,225	3,095	400,845
PEMS-BAY	310,256	58,181	8,192	36,225	3,213	416,067

Table 15: **Mapping from Orion diagram outputs to implementation roles.**

Visible box and output	Implementation role
Input window $X, M : [B, 12, N, 1]$	Model forward call validates and routes the observed signal and validity mask.
Node-local time encoder $C, H^{(0)} : [B, N, D]$ $H^{(\ell)} : [B, N, 3, D]$	Temporal backbone and conditioned token blocks produce local context and lagged states; the training guard applies identity dropout.
Directed influence / lag messages $a : [B, E, 3]$ $M : [B, 12, N, D]$	Dynamic influence weights, graph propagation, and output-residual modules route lagged source messages.
Forecast readout / local readout $Y_{\text{local}} : [B, 12, N, 1]$	Forecast head evaluates horizon-conditioned queries; the output-residual module combines them with the direct local term.
Gated forecast correction $\Delta Y : [B, 12, N, 1]$	Output-residual module applies the graph switch and bounded horizon gate.
Sum $(+)$ $Y_{\text{local}} + \Delta Y$	Output-residual module adds the gated correction to the local forecast.
12-step output $\widehat{Y} : [B, 12, N, 1]$	Model forward call returns the final forecast through its output hook.

The forecast-head depth is set during model construction. The graph's activation epoch and learning-rate ratio are training controls, while node identity, calendar context, width, and the gated correction affect the forward path. PEMS08 selects $W = 128$, $p = .35$, with the graph off. METR-LA and PEMS-BAY select $W = 64$, $p = .10$, with the graph on at epoch 31 and a graph-to-local learning-rate ratio of .10.

H Detailed Orion intervention evidence

Decision phrases below render the accepted machine labels in prose. The original label tokens remain in the supplementary adjudication records.

Table 16 reports the information-premise probe.

Table 16: **Latent-separability probe of the information premise.** The displayed areas under the ROC curve (AUCs) pool all evaluation examples, so their difference is 0.0070. The paired $\Delta\text{AUC} = 0.0169$ is instead the mean of 26 equal-batch AUC differences used by the exact sign-flip rule. Values are rounded for display from the completed validation evidence. The action threshold was set after an earlier point estimate and before this powered rerun.

Hypothesis	A hidden state gives actionable congestion discrimination beyond the scalar forecast.
Frozen criterion	Latent-versus-forecast AUC must pass the paired exact sign-flip test; promotion also requires $\Delta\text{AUC} \geq 0.05$ and AP to pass its paired test.
Result	METR-LA validation: scalar AUC/AP $0.9559/0.7988$; latent AUC/AP $0.9629/0.7393$. Mean paired-batch $\Delta\text{AUC} = 0.0169$, $p_{\text{signflip}} = 0.2582$.
Decision	Inconclusive; the latent state did not meet the action criterion.

Table 17 reports the objective-order intervention.

Table 18 reports the graph-budget intervention.

Table 19 reports the output-head intervention.

Table 17: **Masked-loss objective-order intervention.** Displayed RMSE values are rounded from three-seed validation means. A separate loss-optimum probe is inconclusive. These arms do not establish an optimum.

Hypothesis	Increasing the order of masked error from $p = 1.00$ to 1.25 or 1.50 reduces RMSE without an unacceptable MAE cost.
Frozen criterion	Validation RMSE improvement at least $s_{\max}$, with MAE degradation at most $0.5s_{\max}$; exact $p = 1.00$ replay and all contracted seeds required.
Result	PEMS08 RMSE: control 23.7112; $p = 1.25$, 23.8190; $p = 1.50$, 23.9593. Both changes increased RMSE.
Decision	No RMSE gain for either treatment arm.

Table 18: **Early graph-activation intervention.** Values refer to the earlier-activation arm. The budget-matched schedule study provides context. The separate graph-learning-rate arm is outside this comparison.

Hypothesis	Earlier graph activation gives the graph a longer useful training budget.
Frozen criterion	The earlier-activation arm must improve METR-LA validation MAE over its matched control by at least $s_{\max}$, with all contracted seeds present.
Result	Control MAE 2.874986; earlier-activation MAE 2.874978. The gain was $0.0018s_{\max}$ despite 33–37 graph-training epochs before checkpoint selection.
Decision	Earlier graph activation did not pass the frozen gain rule.

Table 19: **Five-quantile output-head intervention.** The quantile head's parameter change and fixed point-forecast functional are disclosed in the accepted record. Means are rounded for display.

Hypothesis	A $K = 5$ quantile head, converted to a mean point forecast by the frozen reconstruction rule, reduces RMSE.
Frozen criterion	Validation RMSE improvement at least $s_{\max}$; the $K = 1$ median replay must match the original loss exactly.
Result	PEMS08 RMSE: control 23.7112, quantile head 23.8667; improvement $-1.174s_{\max}$. MAE also worsened, by $-1.271s_{\max}$; the head adds 516 parameters (0.075%).
Decision	No RMSE gain.

Table 20 reports the null-exceedance replay.

Table 20: **Zero-effect exceedance replay under the equal-variance normal model.** This post hoc sensitivity replay covered nine treatment arms associated with the earlier 14-hypothesis registry closure. The counts differ because two width arms belong to one hypothesis, while other hypotheses used non-comparable rules or closed after the replay. Within these nine rows, "large-margin" denotes only the three arms with model-based zero-effect crossing probability below 0.002. This descriptive cutoff changes no frozen label and excludes the later P160 and P168 fresh cohorts. The frozen PEMS-BAY comparison changes graph and width jointly, so its small probability does not isolate a width effect.

Frozen gate-passed finding	**Null-exceedance probability**
Graph enabled on PEMS-BAY	4.463e−05
Frozen joint graph–width comparison on PEMS-BAY	6.61e−05
Higher graph learning rate on METR-LA	0.00176
Graph enabled on METR-LA	0.00257
Identity dropout on METR-LA	0.01256
Block-dropout strength 0.35 on PEMS08	0.02409
Higher graph learning rate on PEMS-BAY	0.03005
Width 96 on PEMS08	0.03967
Width 128 on PEMS08	0.05907

I AUDIT-BOUNDARY LEDGER

Table 21 records the accepted audit boundaries.

Table 21: **Accepted audit-boundary snapshot at main-table closure.**

Boundary	Recorded evidence and interpretation
Test exposure	The accepted ledger records 71 logical events: 67 raw test evaluations, 1 attempt aborted before any forward pass, and 3 test-result exposures. Of these, 56 have documented times and 15 have no verified date. Checkpoint/run test passes have a lower bound of 241. Later-use flags are 15 yes, 6 no, and 50 unknown; documented later use informed hypotheses, while the selected recipe components have validation adjudications. The event-level record appears in Appendix F.
Search budget	Orion had 12 recorded validation comparisons on PEMS08, five on METR-LA (including one partial run), and two on PEMS-BAY, with two records missing dataset fields. The PEMS08 count covers comparisons used in the main-table recipe search at closure. Figure 5 instead shows 48 slots across the three datasets from the broader intervention record, including auxiliary, differently ruled, fresh-seed, peer-model, and exploratory displays outside that search count; the omitted width-96 arm appears in Appendix Tables 20 and 27. Baseline tuning counts are unknown. The asymmetric budget may favor Orion; its bias was not estimated.
Hash chronology	One registry digest was a truncated prefix of the current SHA-256. Two contracts were revised before results; 13 contracts without sidecars have no committed Git history. Their self-recorded amendment histories cannot establish an immutable, externally witnessed freeze.
Stopped cases	Eleven runs stopped before a confirmatory conclusion: eight contract defects, one eligibility failure, and two non-evasion guards. Seven have explicit zero validation/test/GPU records; none opened test. Two other stops consumed GPU, so the whole set cannot be called pre-consumption.

Table 22 gives the full-horizon ledger.

Table 22: **Complete test values by forecast horizon.** Each cell is MAE/RMSE/WAPE (%); the 12-step average is repeated for comparison with Table 1. Values are display-rounded from the formal full-precision table.

Model	Horizon	PEMS08	METR-LA	PEMS-BAY
PGCRN	3	12.92/21.35/5.6	2.59/4.96/4.5	1.25/2.62/2.0
PGCRN	6	13.72/23.08/5.9	2.94/6.01/5.1	1.55/3.53/2.5
PGCRN	12	15.01/25.38/6.5	3.36/7.16/5.8	1.83/4.26/2.9
PGCRN	Avg.	13.74/23.05/5.9	2.91/5.99/5.0	1.50/3.45/2.4
HimNet	3	12.68/21.43/5.5	2.61/5.04/4.5	1.27/2.68/2.0
HimNet	6	13.61/23.36/5.9	2.96/6.09/5.1	1.57/3.63/2.5
HimNet	12	14.91/25.76/6.4	3.38/7.25/5.9	1.85/4.36/3.0
HimNet	Avg.	13.56/23.28/5.9	2.93/6.06/5.1	1.52/3.54/2.4
D2STGNN	3	13.37/21.77/5.8	2.57/4.93/4.5	1.26/2.65/2.0
D2STGNN	6	14.46/24.04/6.2	2.92/5.94/5.1	1.57/3.61/2.5
D2STGNN	12	16.08/26.77/7.0	3.36/7.07/5.8	1.87/4.35/3.0
D2STGNN	Avg.	14.43/24.02/6.2	2.89/5.91/5.0	1.52/3.52/2.4
STAEformer	3	12.65/21.55/5.5	2.66/5.11/4.6	1.30/2.77/2.1
STAEformer	6	13.47/23.33/5.8	2.97/6.00/5.1	1.61/3.67/2.6
STAEformer	12	14.98/25.72/6.5	3.34/7.00/5.8	1.87/4.32/3.0
STAEformer	Avg.	13.51/23.30/5.8	2.94/5.97/5.1	1.54/3.56/2.5
Orion	3	12.45/21.31/5.4	2.78/5.49/4.8	1.30/2.78/2.1
Orion	6	13.22/23.06/5.7	3.13/6.54/5.4	1.60/3.71/2.6
Orion	12	14.44/25.15/6.2	3.49/7.45/6.0	1.87/4.37/3.0
Orion	Avg.	13.23/22.98/5.7	3.07/6.42/5.3	1.54/3.59/2.5
AutoSTF-NoNAS†	3	13.35/21.50/5.8	2.70/5.27/4.7	1.29/2.75/2.1
AutoSTF-NoNAS†	6	14.21/23.26/6.1	3.05/6.28/5.3	1.60/3.64/2.6

Model	Horizon	PEMS08	METR-LA	PEMS-BAY (continued)
AutoSTF-NoNAS†	12	15.64/25.68/6.8	3.41/7.25/5.9	1.86/4.30/3.0
AutoSTF-NoNAS†	Avg.	14.23/23.25/6.1	3.00/6.20/5.2	1.53/3.53/2.5
MetaDG	3	12.24/20.85/5.3	3.06/7.25/5.3	1.28/2.72/2.0
MetaDG	6	13.02/22.56/5.6	3.73/9.19/6.5	1.59/3.66/2.5
MetaDG	12	14.21/24.62/6.1	4.63/11.46/8.0	1.86/4.36/3.0
MetaDG	Avg.	13.03/22.47/5.6	3.71/9.24/6.4	1.53/3.56/2.4
ST-SSDL	3	13.64/21.82/5.9	2.60/5.02/4.5	1.32/2.79/2.1
ST-SSDL	6	14.78/23.84/6.4	2.96/6.06/5.1	1.66/3.73/2.7
ST-SSDL	12	16.77/26.76/7.3	3.38/7.21/5.9	1.98/4.50/3.2
ST-SSDL	Avg.	14.86/23.90/6.4	2.92/6.03/5.1	1.60/3.65/2.6
STID	3	13.31/21.56/5.7	2.81/5.53/4.9	1.31/2.78/2.1
STID	6	14.21/23.43/6.1	3.18/6.59/5.5	1.63/3.70/2.6
STID	12	15.57/25.81/6.7	3.54/7.53/6.1	1.90/4.38/3.0
STID	Avg.	14.20/23.39/6.1	3.12/6.48/5.4	1.56/3.60/2.5
Graph-WaveNet	3	13.82/21.87/6.0	2.71/5.19/4.7	1.30/2.75/2.1
Graph-WaveNet	6	14.84/23.77/6.4	3.09/6.22/5.3	1.63/3.71/2.6
Graph-WaveNet	12	16.44/26.21/7.1	3.54/7.30/6.1	1.95/4.51/3.1
Graph-WaveNet	Avg.	14.83/23.71/6.4	3.05/6.17/5.3	1.57/3.63/2.5
TrendGCN	3	14.18/22.35/6.1	2.82/5.46/4.9	1.33/2.82/2.1
TrendGCN	6	15.13/24.20/6.5	3.17/6.43/5.5	1.64/3.76/2.6
TrendGCN	12	16.65/26.62/7.2	3.57/7.40/6.2	1.92/4.45/3.1
TrendGCN	Avg.	15.13/24.15/6.5	3.13/6.35/5.4	1.58/3.65/2.5
AGCRN	3	14.76/23.03/6.4	2.86/5.54/5.0	1.37/2.88/2.2
AGCRN	6	15.90/25.11/6.9	3.24/6.53/5.6	1.70/3.87/2.7
AGCRN	12	17.67/27.91/7.6	3.66/7.51/6.3	1.99/4.59/3.2
AGCRN	Avg.	15.91/25.14/6.9	3.19/6.45/5.5	1.64/3.75/2.6

Table 23 reports seed and rank stability.

Table 23: **Seed variability and rank stability of the complete test ledger.** Each dataset entry is the sample SD across three seeds for 12-step averaged MAE/RMSE/WAPE, with WAPE SD shown in percentage points. The resamples perturb only three observed seeds per model (10,000 draws); rank frequencies are not calibrated probabilities, and the 2.5–97.5% ranges are descriptive, not confidence intervals.

Model	Point rank	Resampled mean-rank range	$P(\text{rank} = 1)$	PEMS08 SD	METR-LA SD	PEMS-BAY SD
PGCRN	2.583	[2.083, 3.417]	98.7%	0.023/0.039/0.010	0.054/0.099/0.094	0.002/0.003/0.004
HimNet	3.889	[3.389, 4.194]	1.7%	0.036/0.114/0.016	0.005/0.019/0.008	0.004/0.011/0.007
D2STGNN	4.222	[3.806, 4.806]	0.0%	0.243/0.148/0.105	0.011/0.044/0.020	0.001/0.008/0.001
STAEformer	4.639	[4.333, 4.917]	0.0%	0.029/0.054/0.012	0.009/0.025/0.016	0.005/0.006/0.008
Orion	5.528	[5.361, 5.861]	0.0%	0.013/0.038/0.006	0.006/0.016/0.010	0.004/0.019/0.006
AutoSTF-NoNAS†	5.556	[5.389, 5.889]	0.0%	0.020/0.016/0.009	0.012/0.029/0.020	0.003/0.008/0.004
MetaDG	5.722	[5.611, 6.028]	0.0%	0.048/0.059/0.021	0.059/0.178/0.103	0.006/0.013/0.009
ST-SSDL	7.889	[6.389, 8.444]	0.0%	0.810/0.677/0.350	0.007/0.026/0.012	0.005/0.010/0.008
STID	8.083	[7.917, 8.389]	0.0%	0.013/0.055/0.006	0.004/0.017/0.007	0.002/0.006/0.004
Graph-WaveNet	8.333	[8.083, 8.611]	0.0%	0.085/0.147/0.037	0.016/0.048/0.028	0.008/0.043/0.012

Model	Point rank	Resampled mean-rank range	$P(\text{rank} = 1)$	PEMS08 SD	METR-LA SD	PEMS-BAY SD (continued)
TrendGCN	10.000	[9.722, 10.111]	0.0%	0.080/0.120/ 0.034	0.006/0.016/ 0.011	0.011/0.029/ 0.017
AGCRN	11.556	[11.472, 11.639]	0.0%	0.231/0.342/ 0.100	0.014/0.053/ 0.024	0.006/0.006/ 0.010

PGCRN is first or tied first in 98.7% of draws. Orion has zero complete-set strict wins in 100.0% of draws; after removing MetaDG, it has 12 strict wins in 75.6%. The number of models gaining a strict-win cell after one-opponent removal has point value 6/12 and a 2.5–97.5% resampling range of [5.0, 7.0]. In the point table, 8/36 top-two gaps are smaller than the larger seed SD.

J Reported-versus-Reproduced Result Ledger

Table 24 contains 186 source-reported cells. Of these, 124 are same-metric comparisons (120 within the 5% numeric threshold, 4 reproduced weaker). The remaining 62 MAPE cells are excluded because the reproduction ledger contains MAE, RMSE, and WAPE, not MAPE. *Source codes:* AG = AGCRN (Bai et al., 2020); SF = STAEformer (Liu et al., 2023a); TE = TESTAM (Lee & Ko, 2024); AU = AutoSTF-NoNAS† (Lyu et al., 2025); MD = MetaDG (Zou et al., 2026); SS = ST-SSDL (Gao et al., 2025); GW = Graph-WaveNet (Wu et al., 2019); SI = STID (Shao et al., 2022a); D2 = D2STGNN (Shao et al., 2022b); HM = HimNet (Dong et al., 2024); TG = TrendGCN (Jiang et al., 2023). T1/T2/T3 identify tables in the cited original paper.

Table 24: **Reported-versus-reproduced result ledger.** Relative deviation is $(\text{reproduced} - \text{reported})/\text{reported}$.

Model	Dataset	Horizon	Metric	Reported	Reproduced	Rel. dev.	Source	Reason
AGCRN	PEMS08	avg12	MAE	15.95	15.9118	-0.24%	AG/T1	In: 5% band
AGCRN	PEMS08	avg12	RMSE	25.22	25.1401	-0.32%	AG/T1	In: 5% band
AGCRN	PEMS08	avg12	MAPE	10.09%	—	—	AG/T1	Out: no MAPE
STAEformer	METR-LA	h3	MAE	2.65	2.6566	+0.25%	SF/T1	In: 5% band
STAEformer	METR-LA	h3	RMSE	5.11	5.1107	+0.01%	SF/T1	In: 5% band
STAEformer	METR-LA	h3	MAPE	6.85%	—	—	SF/T1	Out: no MAPE
STAEformer	METR-LA	h6	MAE	2.97	2.9665	-0.12%	SF/T1	In: 5% band
STAEformer	METR-LA	h6	RMSE	6.00	6.0011	+0.02%	SF/T1	In: 5% band
STAEformer	METR-LA	h6	MAPE	8.13%	—	—	SF/T1	Out: no MAPE
STAEformer	METR-LA	h12	MAE	3.34	3.3407	+0.02%	SF/T1	In: 5% band
STAEformer	METR-LA	h12	RMSE	7.02	7.0045	-0.22%	SF/T1	In: 5% band
STAEformer	METR-LA	h12	MAPE	9.70%	—	—	SF/T1	Out: no MAPE
STAEformer	PEMS-BAY	h3	MAE	1.31	1.3031	-0.52%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h3	RMSE	2.78	2.7732	-0.24%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h3	MAPE	2.76%	—	—	SF/T1	Out: no MAPE
STAEformer	PEMS-BAY	h6	MAE	1.62	1.6057	-0.88%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h6	RMSE	3.68	3.6694	-0.29%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h6	MAPE	3.62%	—	—	SF/T1	Out: no MAPE
STAEformer	PEMS-BAY	h12	MAE	1.88	1.8691	-0.58%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h12	RMSE	4.34	4.3178	-0.51%	SF/T1	In: 5% band
STAEformer	PEMS-BAY	h12	MAPE	4.41%	—	—	SF/T1	Out: no MAPE
STAEformer	PEMS08	avg12	MAE	13.46	13.5111	+0.38%	SF/T3	In: 5% band
STAEformer	PEMS08	avg12	RMSE	23.25	23.3014	+0.22%	SF/T3	In: 5% band
STAEformer	PEMS08	avg12	MAPE	8.88%	—	—	SF/T3	Out: no MAPE
TESTAM	METR-LA	h3	MAE	2.54	2.6460	+4.17%	TE/T1	In: 5% band
TESTAM	METR-LA	h3	RMSE	4.93	5.1390	+4.24%	TE/T1	In: 5% band
TESTAM	METR-LA	h3	MAPE	6.42%	—	—	TE/T1	Out: no MAPE
TESTAM	METR-LA	h6	MAE	2.96	2.9744	+0.49%	TE/T1	In: 5% band
TESTAM	METR-LA	h6	RMSE	6.04	6.0779	+0.63%	TE/T1	In: 5% band
TESTAM	METR-LA	h6	MAPE	7.92%	—	—	TE/T1	Out: no MAPE

Table continued from previous page

Model	Dataset	Horizon	Metric	Reported	Reproduced	Rel. dev.	Source	Reason
TESTAM	METR-LA	h12	MAE	3.36	3.3772	+0.51%	TE/T1	In: 5% band
TESTAM	METR-LA	h12	RMSE	7.09	7.1230	+0.47%	TE/T1	In: 5% band
TESTAM	METR-LA	h12	MAPE	9.67%	—	—	TE/T1	Out: no MAPE
TESTAM	PEMS-BAY	h3	MAE	1.29	1.3001	+0.79%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h3	RMSE	2.77	2.7903	+0.73%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h3	MAPE	2.61%	—	—	TE/T1	Out: no MAPE
TESTAM	PEMS-BAY	h6	MAE	1.59	1.5962	+0.39%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h6	RMSE	3.65	3.6882	+1.05%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h6	MAPE	3.56%	—	—	TE/T1	Out: no MAPE
TESTAM	PEMS-BAY	h12	MAE	1.85	1.8530	+0.16%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h12	RMSE	4.33	4.3199	-0.23%	TE/T1	In: 5% band
TESTAM	PEMS-BAY	h12	MAPE	4.31%	—	—	TE/T1	Out: no MAPE
AutoSTF-NoNAS†	METR-LA	h3	MAE	2.65	2.7030	+2.00%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h3	RMSE	5.10	5.2680	+3.29%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h3	MAPE	6.80%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	METR-LA	h6	MAE	2.99	3.0481	+1.94%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h6	RMSE	6.10	6.2775	+2.91%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h6	MAPE	8.10%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	METR-LA	h12	MAE	3.35	3.4123	+1.86%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h12	RMSE	7.09	7.2495	+2.25%	AU/T2	In: 5% band
AutoSTF-NoNAS†	METR-LA	h12	MAPE	9.58%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	PEMS-BAY	h3	MAE	1.27	1.2941	+1.89%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h3	RMSE	2.70	2.7500	+1.85%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h3	MAPE	2.64%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	PEMS-BAY	h6	MAE	1.56	1.5969	+2.36%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h6	RMSE	3.59	3.6396	+1.38%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h6	MAPE	3.48%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	PEMS-BAY	h12	MAE	1.81	1.8611	+2.82%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h12	RMSE	4.25	4.2966	+1.10%	AU/T2	In: 5% band
AutoSTF-NoNAS†	PEMS-BAY	h12	MAPE	4.23%	—	—	AU/T2	Out: no MAPE
AutoSTF-NoNAS†	PEMS08	avg12	MAE	14.07	14.2342	+1.17%	AU/T3	In: 5% band
AutoSTF-NoNAS†	PEMS08	avg12	RMSE	23.17	23.2549	+0.37%	AU/T3	In: 5% band
AutoSTF-NoNAS†	PEMS08	avg12	MAPE	9.14%	—	—	AU/T3	Out: no MAPE
MetaDG	PEMS08	avg12	MAE	13.04	13.0258	-0.11%	MD/T2	In: 5% band
MetaDG	PEMS08	avg12	RMSE	22.53	22.4732	-0.25%	MD/T2	In: 5% band
MetaDG	PEMS08	avg12	MAPE	8.58%	—	—	MD/T2	Out: no MAPE
ST-SSDL	METR-LA	h3	MAE	2.60	2.6008	+0.03%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h3	RMSE	5.02	5.0208	+0.02%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h3	MAPE	6.62%	—	—	SS/T2	Out: no MAPE
ST-SSDL	METR-LA	h6	MAE	2.96	2.9588	-0.04%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h6	RMSE	6.04	6.0579	+0.30%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h6	MAPE	7.92%	—	—	SS/T2	Out: no MAPE
ST-SSDL	METR-LA	h12	MAE	3.37	3.3827	+0.38%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h12	RMSE	7.17	7.2146	+0.62%	SS/T2	In: 5% band
ST-SSDL	METR-LA	h12	MAPE	9.48%	—	—	SS/T2	Out: no MAPE
ST-SSDL	PEMS-BAY	h3	MAE	1.26	1.3190	+4.68%	SS/T2	In: 5% band
ST-SSDL	PEMS-BAY	h3	RMSE	2.65	2.7856	+5.12%	SS/T2	In: $>5\%$ worse
ST-SSDL	PEMS-BAY	h3	MAPE	2.60%	—	—	SS/T2	Out: no MAPE
ST-SSDL	PEMS-BAY	h6	MAE	1.57	1.6589	+5.66%	SS/T2	In: $>5\%$ worse
ST-SSDL	PEMS-BAY	h6	RMSE	3.59	3.7347	+4.03%	SS/T2	In: 5% band
ST-SSDL	PEMS-BAY	h6	MAPE	3.49%	—	—	SS/T2	Out: no MAPE
ST-SSDL	PEMS-BAY	h12	MAE	1.86	1.9759	+6.23%	SS/T2	In: $>5\%$ worse
ST-SSDL	PEMS-BAY	h12	RMSE	4.34	4.4981	+3.64%	SS/T2	In: 5% band
ST-SSDL	PEMS-BAY	h12	MAPE	4.35%	—	—	SS/T2	Out: no MAPE

Table continued from previous page

Model	Dataset	Horizon	Metric	Reported	Reproduced	Rel. dev.	Source	Reason
ST-SSDL	PEMS08	avg12	MAE	13.86	14.8647	+7.25%	SS/T2	In: $> 5\%$ worse
ST-SSDL	PEMS08	avg12	RMSE	23.10	23.8979	+3.45%	SS/T2	In: 5% band
ST-SSDL	PEMS08	avg12	MAPE	9.17%	—	—	SS/T2	Out: no MAPE
Graph-WaveNet	METR-LA	h3	MAE	2.69	2.7061	+0.60%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h3	RMSE	5.15	5.1855	+0.69%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h3	MAPE	6.90%	—	—	GW/T2	Out: no MAPE
Graph-WaveNet	METR-LA	h6	MAE	3.07	3.0866	+0.54%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h6	RMSE	6.22	6.2162	-0.06%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h6	MAPE	8.37%	—	—	GW/T2	Out: no MAPE
Graph-WaveNet	METR-LA	h12	MAE	3.53	3.5379	+0.22%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h12	RMSE	7.37	7.3045	-0.89%	GW/T2	In: 5% band
Graph-WaveNet	METR-LA	h12	MAPE	10.01%	—	—	GW/T2	Out: no MAPE
Graph-WaveNet	PEMS-BAY	h3	MAE	1.30	1.3049	+0.38%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h3	RMSE	2.74	2.7504	+0.38%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h3	MAPE	2.73%	—	—	GW/T2	Out: no MAPE
Graph-WaveNet	PEMS-BAY	h6	MAE	1.63	1.6337	+0.23%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h6	RMSE	3.70	3.7138	+0.37%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h6	MAPE	3.67%	—	—	GW/T2	Out: no MAPE
Graph-WaveNet	PEMS-BAY	h12	MAE	1.95	1.9461	-0.20%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h12	RMSE	4.52	4.5053	-0.33%	GW/T2	In: 5% band
Graph-WaveNet	PEMS-BAY	h12	MAPE	4.63%	—	—	GW/T2	Out: no MAPE
STID	PEMS08	h3	MAE	13.28	13.3078	+0.21%	SI/T3	In: 5% band
STID	PEMS08	h3	RMSE	21.66	21.5571	-0.48%	SI/T3	In: 5% band
STID	PEMS08	h3	MAPE	8.62%	—	—	SI/T3	Out: no MAPE
STID	PEMS08	h6	MAE	14.21	14.2098	-0.00%	SI/T3	In: 5% band
STID	PEMS08	h6	RMSE	23.57	23.4300	-0.59%	SI/T3	In: 5% band
STID	PEMS08	h6	MAPE	9.24%	—	—	SI/T3	Out: no MAPE
STID	PEMS08	h12	MAE	15.58	15.5671	-0.08%	SI/T3	In: 5% band
STID	PEMS08	h12	RMSE	25.89	25.8140	-0.29%	SI/T3	In: 5% band
STID	PEMS08	h12	MAPE	10.33%	—	—	SI/T3	Out: no MAPE
STID	PEMS08	avg12	MAE	14.20	14.2004	+0.00%	SI/T3	In: 5% band
STID	PEMS08	avg12	RMSE	23.49	23.3883	-0.43%	SI/T3	In: 5% band
STID	PEMS08	avg12	MAPE	9.28%	—	—	SI/T3	Out: no MAPE
STID	PEMS-BAY	h3	MAE	1.30	1.3082	+0.63%	SI/T3	In: 5% band
STID	PEMS-BAY	h3	RMSE	2.81	2.7777	-1.15%	SI/T3	In: 5% band
STID	PEMS-BAY	h3	MAPE	2.73%	—	—	SI/T3	Out: no MAPE
STID	PEMS-BAY	h6	MAE	1.62	1.6267	+0.41%	SI/T3	In: 5% band
STID	PEMS-BAY	h6	RMSE	3.72	3.7044	-0.42%	SI/T3	In: 5% band
STID	PEMS-BAY	h6	MAPE	3.68%	—	—	SI/T3	Out: no MAPE
STID	PEMS-BAY	h12	MAE	1.89	1.9029	+0.68%	SI/T3	In: 5% band
STID	PEMS-BAY	h12	RMSE	4.40	4.3838	-0.37%	SI/T3	In: 5% band
STID	PEMS-BAY	h12	MAPE	4.47%	—	—	SI/T3	Out: no MAPE
STID	PEMS-BAY	avg12	MAE	1.55	1.5605	+0.68%	SI/T3	In: 5% band
STID	PEMS-BAY	avg12	RMSE	3.62	3.5953	-0.68%	SI/T3	In: 5% band
STID	PEMS-BAY	avg12	MAPE	3.51%	—	—	SI/T3	Out: no MAPE
D2STGNN	METR-LA	h3	MAE	2.56*	2.5712	+0.44%	D2/T3	In: 5% band
D2STGNN	METR-LA	h3	RMSE	4.88*	4.9345	+1.12%	D2/T3	In: 5% band
D2STGNN	METR-LA	h3	MAPE	6.48%*	—	—	D2/T3	Out: no MAPE
D2STGNN	METR-LA	h6	MAE	2.90*	2.9168	+0.58%	D2/T3	In: 5% band
D2STGNN	METR-LA	h6	RMSE	5.89*	5.9371	+0.80%	D2/T3	In: 5% band
D2STGNN	METR-LA	h6	MAPE	7.78%*	—	—	D2/T3	Out: no MAPE
D2STGNN	METR-LA	h12	MAE	3.35*	3.3627	+0.38%	D2/T3	In: 5% band
D2STGNN	METR-LA	h12	RMSE	7.03*	7.0683	+0.54%	D2/T3	In: 5% band
D2STGNN	METR-LA	h12	MAPE	9.40%*	—	—	D2/T3	Out: no MAPE
D2STGNN	PEMS-BAY	h3	MAE	1.24*	1.2562	+1.31%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h3	RMSE	2.60*	2.6451	+1.74%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h3	MAPE	2.58%*	—	—	D2/T3	Out: no MAPE

Table continued from previous page

Model	Dataset	Horizon	Metric	Reported	Reproduced	Rel. dev.	Source	Reason
D2STGNN	PEMS-BAY	h6	MAE	1.55*	1.5727	+1.46%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h6	RMSE	3.52*	3.6053	+2.42%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h6	MAPE	3.49%*	—	—	D2/T3	Out: no MAPE
D2STGNN	PEMS-BAY	h12	MAE	1.85*	1.8669	+0.91%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h12	RMSE	4.30*	4.3477	+1.11%	D2/T3	In: 5% band
D2STGNN	PEMS-BAY	h12	MAPE	4.37%	—	—	D2/T3	Out: no MAPE
D2STGNN	PEMS08	h3	MAE	13.14*	13.3670	+1.73%	D2/T3	In: 5% band
D2STGNN	PEMS08	h3	RMSE	21.42*	21.7694	+1.63%	D2/T3	In: 5% band
D2STGNN	PEMS08	h3	MAPE	8.55%*	—	—	D2/T3	Out: no MAPE
D2STGNN	PEMS08	h6	MAE	14.21*	14.4630	+1.78%	D2/T3	In: 5% band
D2STGNN	PEMS08	h6	RMSE	23.65*	24.0407	+1.65%	D2/T3	In: 5% band
D2STGNN	PEMS08	h6	MAPE	9.12%*	—	—	D2/T3	Out: no MAPE
D2STGNN	PEMS08	h12	MAE	15.69*	16.0818	+2.50%	D2/T3	In: 5% band
D2STGNN	PEMS08	h12	RMSE	26.41	26.7743	+1.38%	D2/T3	In: 5% band
D2STGNN	PEMS08	h12	MAPE	10.17%*	—	—	D2/T3	Out: no MAPE
HimNet	METR-LA	h3	MAE	2.60	2.6096	+0.37%	HM/T3	In: 5% band
HimNet	METR-LA	h3	RMSE	5.02	5.0432	+0.46%	HM/T3	In: 5% band
HimNet	METR-LA	h3	MAPE	6.70%	—	—	HM/T3	Out: no MAPE
HimNet	METR-LA	h6	MAE	2.95	2.9617	+0.40%	HM/T3	In: 5% band
HimNet	METR-LA	h6	RMSE	6.06	6.0869	+0.44%	HM/T3	In: 5% band
HimNet	METR-LA	h6	MAPE	8.11%	—	—	HM/T3	Out: no MAPE
HimNet	METR-LA	h12	MAE	3.37	3.3809	+0.32%	HM/T3	In: 5% band
HimNet	METR-LA	h12	RMSE	7.22	7.2487	+0.40%	HM/T3	In: 5% band
HimNet	METR-LA	h12	MAPE	9.79%	—	—	HM/T3	Out: no MAPE
HimNet	PEMS-BAY	h3	MAE	1.27	1.2691	-0.07%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h3	RMSE	2.68	2.6849	+0.18%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h3	MAPE	2.64%	—	—	HM/T3	Out: no MAPE
HimNet	PEMS-BAY	h6	MAE	1.57	1.5744	+0.28%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h6	RMSE	3.60	3.6294	+0.82%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h6	MAPE	3.52%	—	—	HM/T3	Out: no MAPE
HimNet	PEMS-BAY	h12	MAE	1.84	1.8534	+0.73%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h12	RMSE	4.32	4.3610	+0.95%	HM/T3	In: 5% band
HimNet	PEMS-BAY	h12	MAPE	4.33%	—	—	HM/T3	Out: no MAPE
HimNet	PEMS08	avg12	MAE	13.57	13.5602	-0.07%	HM/T3	In: 5% band
HimNet	PEMS08	avg12	RMSE	23.22	23.2807	+0.26%	HM/T3	In: 5% band
HimNet	PEMS08	avg12	MAPE	8.98%	—	—	HM/T3	Out: no MAPE
TrendGCN	METR-LA	h12	MAE	3.55	3.5719	+0.62%	TG/T3	In: 5% band
TrendGCN	METR-LA	h12	RMSE	7.39	7.4023	+0.17%	TG/T3	In: 5% band
TrendGCN	METR-LA	h12	MAPE	10.27%	—	—	TG/T3	Out: no MAPE
TrendGCN	PEMS-BAY	h12	MAE	1.92	1.9170	-0.16%	TG/T3	In: 5% band
TrendGCN	PEMS-BAY	h12	RMSE	4.46	4.4547	-0.12%	TG/T3	In: 5% band
TrendGCN	PEMS-BAY	h12	MAPE	4.51%	—	—	TG/T3	Out: no MAPE
TrendGCN	PEMS08	avg12	MAE	15.15	15.1312	-0.12%	TG/T2	In: 5% band
TrendGCN	PEMS08	avg12	RMSE	24.26	24.1539	-0.44%	TG/T2	In: 5% band
TrendGCN	PEMS08	avg12	MAPE	9.51%	—	—	TG/T2	Out: no MAPE

Ledger scope. "Included" means a same-metric numerical comparison, even when the original protocol differs. It does not certify methodological equivalence. avg12 is the 12-step micro-average. PGCRN contributes zero rows to the 124-comparison denominator because its original full text is unverified and no source-confirmed numerical cells were extracted. It remains an eligible forecasting comparator. A star in a reported cell reproduces the original paper's mark. It is not a significance claim from this ledger.

K Additional Audit Figures

Figures 10–13 report horizon-wise test errors, comparator-removal sensitivity, reported-versus-reproduced agreement, and cumulative test exposure.

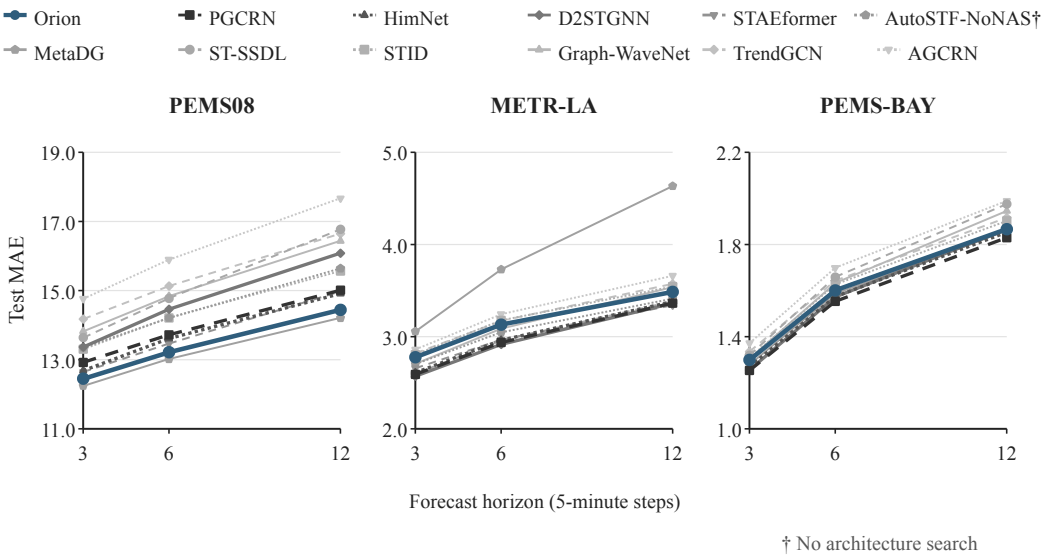

Figure 10: Test mean absolute error (MAE) at horizons 3, 6, and 12 for all complete-coverage models across PEMS08 (PeMS District 8 flow), METR-LA (Los Angeles speed), and PEMS-BAY (Bay Area speed). Each dataset has its own labeled vertical scale. Connecting lines link observed horizons only. The dagger marks AutoSTF-NoNAS, the variant without architecture search.

Figure 14 summarizes the diagnostic-to-repair chain and the frozen outcomes of one observational probe and three intervention tests.

L Selected Validation Cases and Error Contributions

Figures 15–16 report the selected validation cases and the corresponding signed error contributions.

M Additional evidence added before submission

M.1 Bitcoin anchoring of timestamp receipts

This receipt audit adds no registered hypothesis and addresses whether the reported public timestamp submissions acquired Bitcoin confirmation before the corresponding results. The audit ran `ots upgrade` on the located receipts, extracted the Bitcoin attestations, and compared each proof's Merkle-root bytes (after reversing the proof byte order) with the root in the public block header. A mapped receipt commitment is included in a Bitcoin block whose public header reports the stated time. The header time is an approximate timestamp for the containing block, not a strict bound on the moment when the proof began to exist. Bitcoin header times can differ from true mining times by up to about two hours, so an anchor is decisive only when it precedes the first result by more than that tolerance. Native `ots verify` with local-chain verification was unavailable because this machine had no Bitcoin Core node. The reported root and header comparisons used public Blockstream records.

Of the 48 registered contracts, 34 have a mapped Bitcoin confirmation. Eight of the 48 contracts carry Bitcoin anchors whose block times precede their first results or no-run bounds; five of them do so by more than the two-hour tolerance. All eight previously reported public submissions are in this group. For 22, the block time follows the first verifiable result and therefore cannot establish a pre-result freeze. Four confirmed contracts lack a strict result-time comparison, and 14 have no mapped

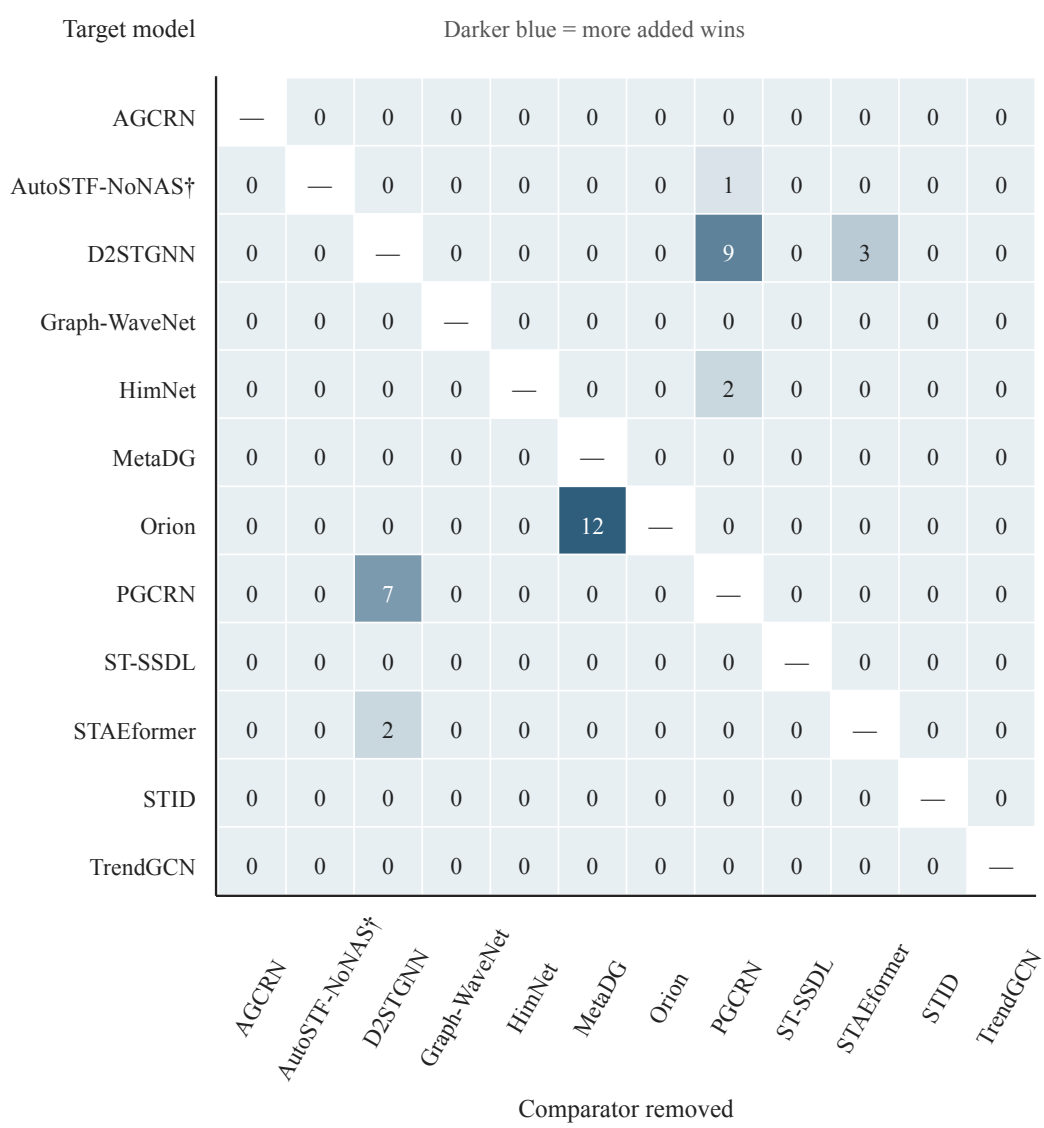

Figure 11: **Sensitivity of strictly best test-cell counts to removing one comparator.** Each cell gives the additional wins for the row model after removing the column model from the same 36-cell universe. The diagonal is inapplicable. Measured zero changes remain visible. The dagger marks AutoSTF-NoNAS, the variant without architecture search.

Bitcoin confirmation. These categories exhaust the 48 contracts. In the supplementary package, every receipt either matches its adjacent file or is labeled as attesting to the unsanitized original. The three contracts added before submission (P169–P171), which lie outside the 48, were later anchored in Bitcoin block 968554 (2026-09-25 13:55:15 UTC). This block time precedes the first P169 result (14:02:50 UTC) and the first P171 result. It follows the first P170 result (13:38:53 UTC). The P170 public-calendar submission (12:48:38 UTC) preceded all of its runs.

Table 25 lists the eight header-before comparisons.

M.2 Post hoc multiplicity and paired-seed summaries

This analysis adds no registered hypothesis and addresses multiplicity and the sensitivity of paired-seed summaries to three-seed sampling. It is post hoc and descriptive. Neither adjustment nor paired analysis changed a frozen decision rule or label. For the family of 45 one-sided decisions, 12 crossed the original support gate. The null-exceedance probabilities were computed from the

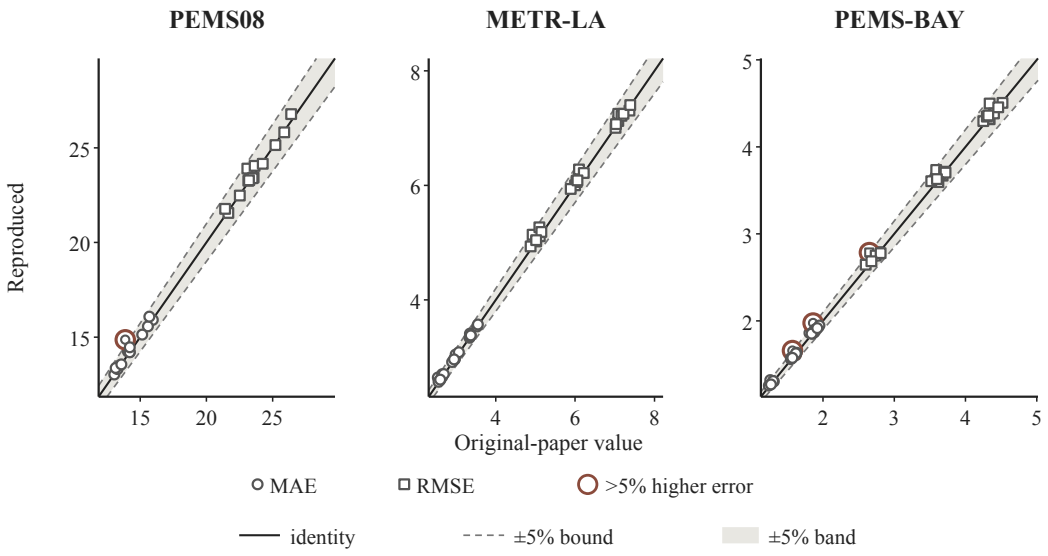

Figure 12: Reproduced test errors against original-paper reports for comparable mean absolute error (MAE) and root mean squared error (RMSE) records. Panels show PEMS08 (PeMS District 8 flow), METR-LA (Los Angeles speed), and PEMS-BAY (Bay Area speed). Each panel uses equal horizontal and vertical numerical scales. The solid line is identity. Dashed lines bound a $\pm 5\%$ relative band. Rust rings mark records whose reproduced error exceeds the report by more than 5%. Reported mean absolute percentage error has no like-for-like reproduced value and is excluded.

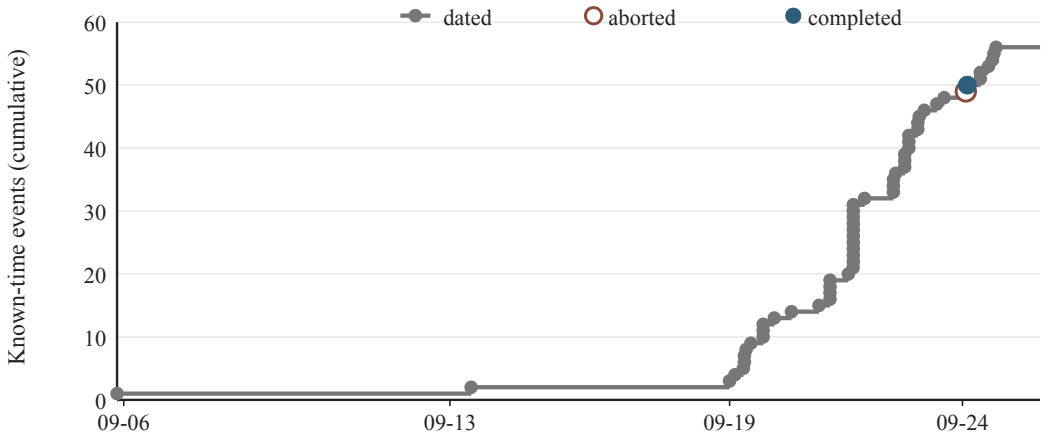

Figure 13: **Cumulative logical test-exposure events in UTC.** The curve contains all 56 entries with documented times in the final 71-event ledger; the 15 events without verified times are counted separately and excluded from the chronology. After UTC sorting, the September 24 PEMS08 (PeMS District 8) opening that aborted before yielding a numeric result is event 49, and the completed evaluation is event 50. The ledger includes baseline and Orion events.

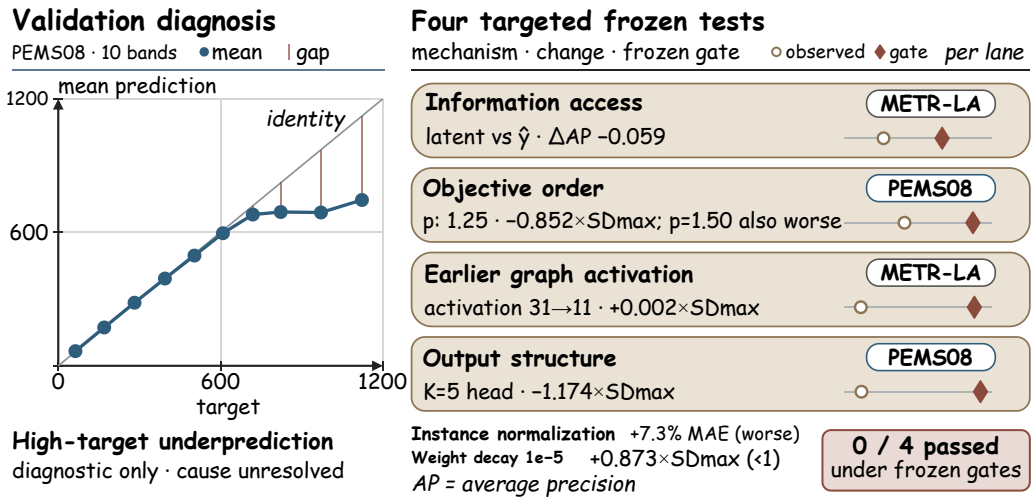

Figure 14: **Diagnostic-to-repair chain.** The left panel plots ten PEMS08 validation-band means. It is a band-mean diagnostic, not a raw-sample trajectory. Because the diagnostic bins observations by realized targets, underprediction in bins with high realized targets alone does not establish forecast miscalibration or a correctable modeling defect. The pattern motivated one observational probe and three targeted intervention tests, shown with their validation change (open circle) and frozen acceptance gate (diamond) on lane-local scales. Rightward means larger normalized gain within a lane. Marker positions are not comparable across lanes. Filled circles denote band means, and vertical rust segments show their gaps to the identity line. AP denotes average precision. In-figure SDmax denotes $s_{\max}$, the larger arm-level standard deviation. The METR-LA graph-activation test moved activation from epoch 31 to 11, increasing graph-training exposure from 12–18 to 33–37 epochs. Its end-to-end gain was only $+0.002\, s_{\max}$. None of the four evaluations met its frozen action criterion under the tested settings. In the footer, positive normalized gain means lower MAE: the best weight-decay arm (10^{-5}) reached $+0.873\, s_{\max}$, below the $1\times$ gate. The instance-normalization value instead reports relative MAE increase, so $+7.3\%$ means worse. These outcomes leave the cause of the high-target underprediction unresolved and do not identify a mechanism. The objective-order lane shows the $p = 1.25$ margin; the $p = 1.50$ arm also increased RMSE (Table 17).

Table 25: **Bitcoin block-header times for eight contracts whose headers precede their first results or no-run bounds; five exceed the two-hour tolerance.** "No-run bound" is a documented time by which no formal run had occurred. It is not a result timestamp. Times are UTC.

Contract	Height	Block time (UTC)	First verified event (UTC)
P148: PEMS08 dropout sweep	968268	2026-09-23 11:44:34	2026-09-23 16:22:27.419587
P149: PEMS08 weight decay and main rate	968268	2026-09-23 11:44:34	2026-09-23 17:29:19
P156/P157: PEMS08 gate and fraction calibration	968268	2026-09-23 11:44:34	2026-09-24 01:19:20[a]
P160: PEMS08 dropout replication	968284	2026-09-23 14:38:16	2026-09-24 01:19:20[a]
P161: external-model default audit	968284	2026-09-23 14:38:16	no-run bound: 2026-09-23 14:39:38
P162: PEMS08 unscanned defaults	968284	2026-09-23 14:38:16	no-run bound: 2026-09-23 17:07:53
P165: PEMS08 real A/A calibration	968379	2026-09-24 09:36:20	2026-09-24 09:37:32[b]
P168: METR-LA graph rate; PEMS-BAY graph on	968405	2026-09-24 14:43:55	2026-09-24 14:48:03[c]

[a] Combined package first-start lower bound. [b] First queue launch. [c] First GPU-process start. The first two rows use recorded result completions. The margin exceeds two hours for P148, P149, P156/P157, P160, and P162; it does not for P161, P165, or P168.

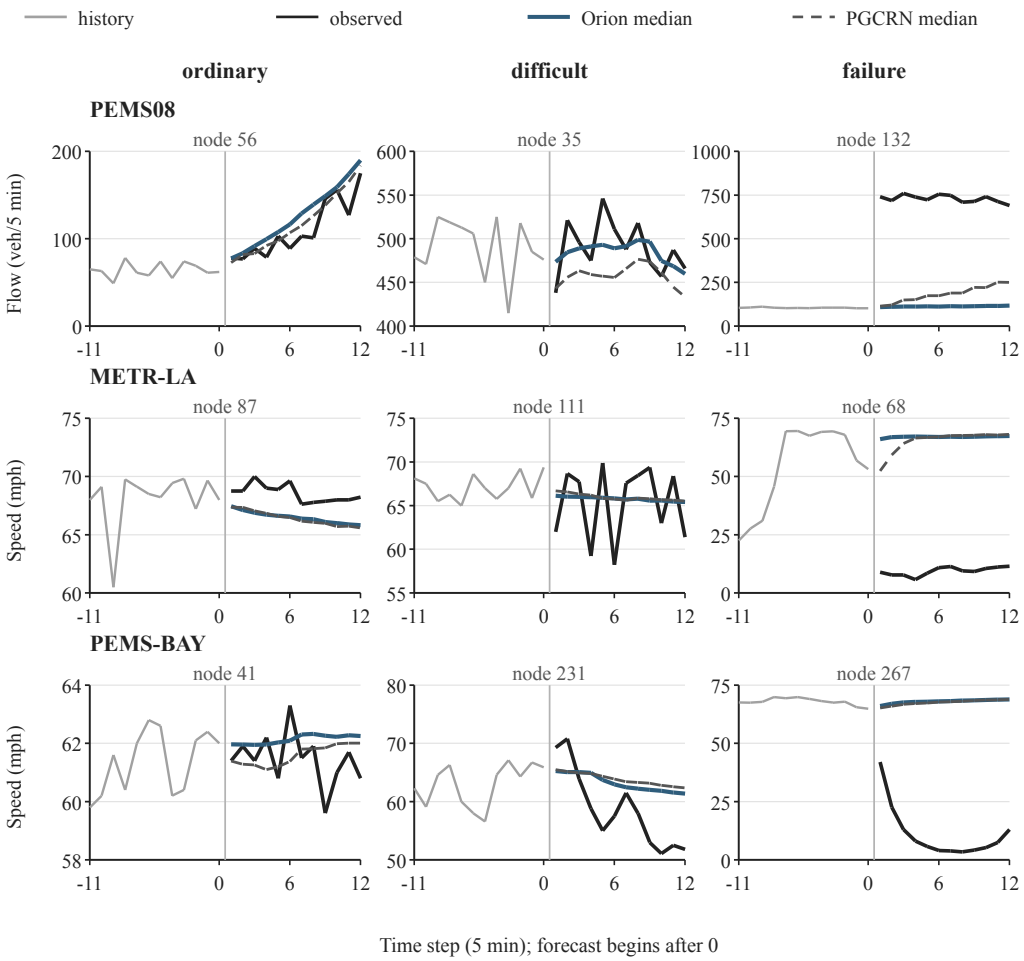

Figure 15: Nine cases from the selection validation split (used for recipe selection, not held out). The frozen selector chooses ordinary and difficult cases from target-derived difficulty, then a failure case from the largest remaining Orion self-error. It fixes all cases before consulting PGCRN. Each line is an observed series or the elementwise median forecast from three fixed seeds. PEMS08 (PeMS District 8) records five-minute traffic flow in vehicles. METR-LA (Los Angeles) and PEMS-BAY (Bay Area) record traffic speed in miles per hour. The observed values change abruptly in the failure cases, and both forecasts miss those changes. This observation does not establish a cause of the errors.

observed standardized margins under the equal-variance normal null, then adjusted across the 45 decisions by Benjamini–Hochberg (BH) and Holm. A decision appears below only if it both crossed the original gate and met the stated adjusted threshold. Some decisions reuse a control triplet. These lists are descriptive and do not assert a family-level error-rate guarantee. Holm's procedure does not require independent p-values; we make no family-wise claim because the model-based tail probabilities have not been established as valid p-values for this adaptive analysis.

Table 26 lists the comparisons flagged by the post hoc descriptive thresholds.

For paired summaries, $d_i = C_i - T_i$ uses equal seed identifiers in ascending order. The displayed interval is the descriptive t_2 interval $\bar{d} \pm t_{0.975,2} s_d/\sqrt{3}$. Fifteen of 18 requested rows admit three paired differences. The three unavailable rows are the single-seed permutation-bound case, an identity-shrinkage comparison whose treatment values were not retained per seed in the local authority record, and the 18-seed A/A calibration, which has no single three-pair control–treatment contrast. No missing differences were imputed.

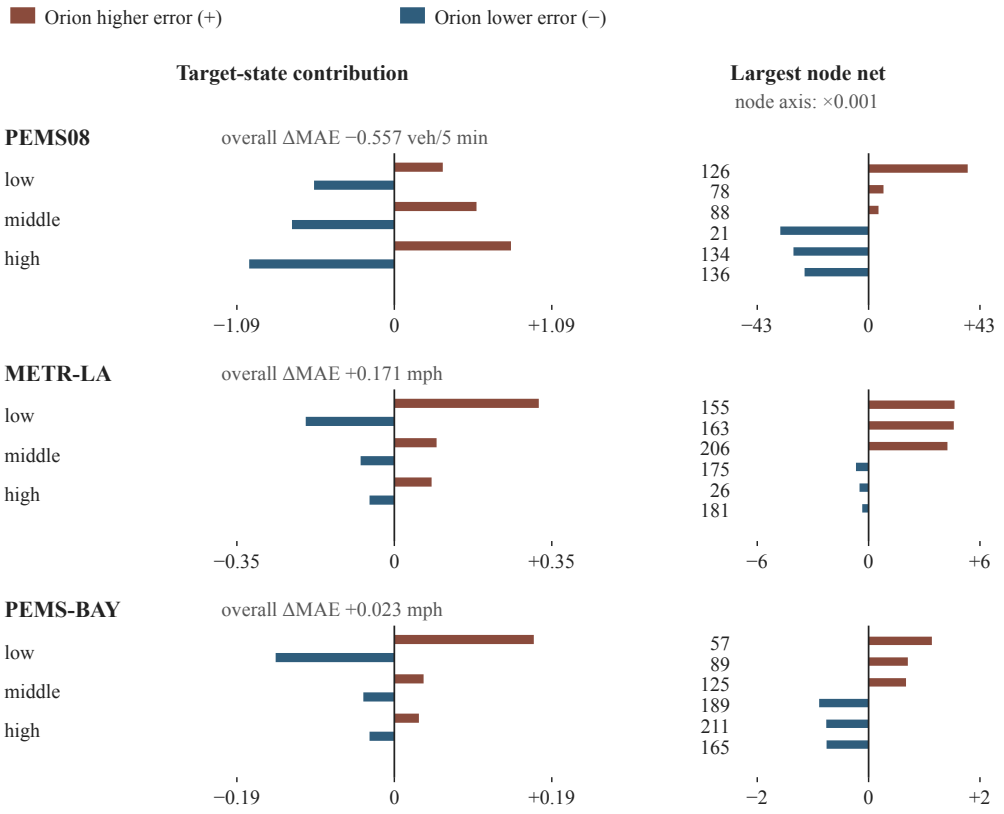

Figure 16: Signed contributions to the Orion–PGCRN mean absolute error (MAE) gap on the selection validation split (used for recipe selection, not held out). Low, middle, and high states use tertiles of valid training targets within each dataset. The positive and negative bars divide signed error sums by all valid validation targets, so their sum equals the displayed overall gap. Positive means higher Orion error. Node-axis ticks are in units of 10^{-3} MAE. The lowest speed tertile accounts for most of the net positive gap on METR-LA (Los Angeles) and PEMS-BAY (Bay Area), whereas PEMS08 (PeMS District 8) has a negative overall gap on this split. These validation observations make no claim about the test split.

Table 26: **Post hoc BH/Holm thresholds applied to model-based, control-dependent zero-effect tail probabilities.** Counts are descriptive flags, not discoveries controlled for the false discovery rate or family-wise error rate. Names are the neutral comparison names used in the appendix registry.

Adjustment	Level	Count	Flagged comparisons
BH	$q = 0.05$	6	METR-LA graph activation; PEMS-BAY graph activation; METR-LA higher graph learning rate; PEMS-BAY width 64 to 128 with graph; PEMS08 block dropout 0.35, fresh seeds; PEMS-BAY fresh graph on.
BH	$q = 0.10$	8	The preceding six; METR-LA identity block dropout; METR-LA fresh graph rate 0.10.
BH	$q = 0.20$	11	The preceding eight; PEMS08 width 96; PEMS-BAY higher graph learning rate; PEMS08 block dropout 0.35.
Holm	$\alpha = 0.05$	4	PEMS-BAY graph activation; PEMS-BAY width 64 to 128 with graph; PEMS08 block dropout 0.35, fresh seeds; PEMS-BAY fresh graph on.
Holm	$\alpha = 0.10$	5	The preceding four; METR-LA higher graph learning rate.
Holm	$\alpha = 0.20$	6	The preceding five; METR-LA graph activation.

Table 27 reports the paired-seed summaries.

Table 27: **Descriptive paired validation-MAE differences, $C - T$.** Values are shown at the precision of the accepted evidence report; unavailable rows retain their recorded reason.

Comparison	$\bar{d}$	s_d	t_2 95% interval / reason
PEMS08 gap-location bound	–	–	One matched seed; no three-pair gate
PEMS08 graph activation	-0.032504	0.027258	[-0.100217, 0.035209]
PEMS08 deterministic identity shrinkage	–	–	Treatment triplet absent locally
METR-LA identity block dropout	0.013052	0.004045	[0.003005, 0.0231]
METR-LA graph activation	0.020832	0.001257	[0.01771, 0.023954]
PEMS-BAY graph activation	0.017619	0.0005883	[0.016158, 0.01908]
PEMS08 width 96	0.049229	0.040314	[-0.050917, 0.149375]
PEMS08 width 128	0.061586	0.064158	[-0.097792, 0.220964]
METR-LA higher graph learning rate	0.018004	0.0007350	[0.016178, 0.01983]
PEMS-BAY width 64 to 128 with graph	0.020104	0.001637	[0.016038, 0.02417]
PEMS08 graph learning rate 0.10	-0.066491	0.008530	[-0.087681, -0.045301]
PEMS-BAY higher graph learning rate	0.002631	0.001014	[0.0001114, 0.005151]
PEMS08 block dropout 0.35	0.087634	0.038865	[-0.008912, 0.18418]
PEMS08 batch size 32	-0.061677	0.031531	[-0.140004, 0.016649]
PEMS08 real A/A calibration	–	–	No single three-pair contrast
PEMS08 block dropout 0.35, fresh seeds	0.130430	0.015529	[0.091854, 0.169006]
METR-LA fresh graph rate 0.10	0.015050	0.005107	[0.002363, 0.027737]
PEMS-BAY fresh graph on	0.019026	0.002743	[0.012212, 0.025841]

Of the 15 computable intervals, ten exclude zero and five include it. For a normal paired difference with $n = 3$, the sample standard deviation has relative sampling coefficient of variation $\sqrt{4/\pi - 1} = 52.27\%$. The corresponding population-standard-deviation interval spans $[0.5207 s_d, 6.2847 s_d]$ at 95% confidence. This uncertainty limits interpretation of individual paired intervals and does not revise the frozen gate.

M.3 A POSITIVE CONTROL

This protocol-level calibration is outside the 48 registered hypotheses and asks whether the frozen gate recovers two previously measured harms when they are reversed on fresh PEMS08 validation seeds. The harmful adaptive instance-normalization and static-normalization configurations serve separately as controls. The unchanged base serves as the arm. The three fresh seeds are 3425–3427. Each of the nine runs uses the frozen PEMS08 configuration and 150 epochs. For each comparison, the predefined gain is mean(control) minus mean(base), and the original gate requires it to be at least the larger sample standard deviation of the two three-seed arms. The two previously measured harms specify direction and an approximate magnitude before these fresh-seed runs.

Table 28 lists the fresh-seed values.

Table 28: **Positive-control validation MAE by arm and fresh seed.** The two normalization arms are controls. BASE is the common comparison arm.

Arm	Seed 3425	Seed 3426	Seed 3427
BASE	13.351496275176808	13.497984947060528	13.407449693185006
NORM_A	14.332473438742182	14.351313022362390	14.353409414105460
NORM_B	13.629583123074475	13.660198797409473	13.492928438215928

Both reversals crossed the frozen gate. The adaptive-normalization reversal had gain 0.926755 MAE, $s_{\max} = 0.073922$, and $m = \text{gain}/s_{\max} = 12.537012$. The static-normalization reversal had gain 0.175260, $s_{\max} = 0.089061$, and $m = 1.967862$. Each gain lies inside its prespecified band around the earlier P150 estimate, so both decisions are `RECOVERED` and the contracted overall label is `POSITIVE_CONTROL_RECOVERED`. These are two known-harm reversals on one dataset. They do not measure power for smaller effects.

Table 29 reports the positive-control decisions.

Table 29: **Frozen positive-control adjudication.** The prior band is $[0.5, 1.5]$ times the earlier P150 gain and is descriptive. Displayed summaries are rounded for space. Labels use the full-precision values in the accepted adjudication.

Decision	Control	Gain	$s_{\max}$	m	Prior gain	Prior band	Label
PC_A	NORM_A	0.926755	0.073922	12.537012	0.980647	[0.490324, 1.470971]	RECOVERED
PC_B	NORM_B	0.175260	0.089061	1.967862	0.253588	[0.126794, 0.380381]	RECOVERED

The adjudicator passed 279 of 279 identity and discipline checks: every configuration difference after removing the seed was empty, all runs completed, and no run opened the test split. The P169 contract was submitted to public OpenTimestamps calendars at 12:42 UTC. The adjudicator followed at 13:00:33 UTC, and the first result-producing run began at 13:08:43.980108 UTC. A wrapper attempt at 13:06:21 UTC stopped before epoch 0 and produced no validation result. These calendar-submission times establish pre-run order. A later receipt upgrade anchored the P169 contract in Bitcoin block 968554, whose block time preceded the first P169 result.

M.4 A SECOND MODEL UNDER THE SAME PROTOCOL

This second-model case is outside Orion's 48 registered hypotheses and asks whether the same prospectively frozen procedure can evaluate three embedding components of the published STID implementation. The accepted full STID configuration provides the three-seed reference on PEMS08 and METR-LA. Each ablation turns off exactly one of the spatial-identity, time-of-day, or day-of-week embeddings. The remaining settings follow the accepted implementation. Before ablations, a full-model replay through the ablation launcher must match the accepted validation MAE bit-exactly for seed 3407 on each dataset. Otherwise that dataset receives no verdict. The six component comparisons use seeds 3407–3409, 100 epochs, validation micro-MAE, and the same mean-gain versus maximum sample-SD gate. The test split was not opened for these comparisons.

All three removals used existing STID configuration switches. The model source was not edited. Table 30 gives the parameter counts recorded for the accepted FULL configuration and each single-component ablation.

Table 30: **STID component switches and parameter counts.** Each NO arm changes one switch from the accepted FULL configuration. SID, TID, and DIW denote the spatial-identity, time-of-day, and day-of-week embeddings.

Arm	Switch set to false	PEMS08 parameters	METR-LA parameters
FULL	none	116684	117868
NO_SID	`if_node`	67660	67660
NO_TID	`if_T_i_D`	63884	65068
NO_DIW	`if_D_i_W`	72876	74060

Before the ablations, the same launcher reproduced the accepted FULL seed-3407 validation value bit-exactly on both datasets: 14.457868527788628 on PEMS08 and 2.9258836821760386 on METR-LA. The accepted and replayed values were equal in each case, with a recorded difference of zero and no test opening.

Table 31 reports the component decisions.

Four component comparisons passed the gate, one was flat, and one was reversed. These labels describe this implementation and validation protocol. They are not a test-set result or a claim that the original STID paper is incorrect.

Table 32 lists the component values by seed.

The contract was submitted to public OpenTimestamps calendars at 12:48:38 UTC. The construction checks ran before adjudicator freeze. The adjudicator was submitted at 13:26:50 UTC, and a prelaunch receipt was written at 13:27:08 UTC. The first formal ablation began at 13:29:45.019991

Table 31: **STID component decisions on validation micro-MAE.** Control is the component-removed arm. FULL is the accepted reference. Here $m = \text{gain}/s_{\max}$, not gain minus $s_{\max}$. Displayed numbers are rounded for space. Adjudication used full precision.

Dataset	Removed	Control mean	FULL mean	Gain	$s_{\max}$	m	Label
PEMS08	SID	16.233039	14.452756	1.780283	0.027147	65.579699	GATE_PASSED
PEMS08	TID	15.110407	14.452756	0.657651	0.027147	24.225680	GATE_PASSED
PEMS08	DIW	14.477522	14.452756	0.024766	0.027147	0.912310	FLAT
METR-LA	SID	3.198397	2.922278	0.276119	0.003624	76.186861	GATE_PASSED
METR-LA	TID	2.951721	2.922278	0.029443	0.003673	8.015518	GATE_PASSED
METR-LA	DIW	2.914246	2.922278	-0.008031	0.005492	-1.462334	REVERSED

Table 32: **STID validation micro-MAE by dataset, arm, and seed.** FULL values come from the accepted baseline records. Each NO arm was run under the frozen second-model contract.

Dataset	Arm	Seed 3407	Seed 3408	Seed 3409
PEMS08	FULL	14.457868527788628	14.423415880648683	14.476982513951274
PEMS08	NO_SID	16.220694196744603	16.245571745660904	16.232850608117843
PEMS08	NO_TID	15.120481939301510	15.114179460389450	15.096559308806160
PEMS08	NO_DIW	14.469530527506018	14.472113920710258	14.490921566824737
METR-LA	FULL	2.9258836821760386	2.9223146097229180	2.9186354870583210
METR-LA	NO_SID	3.1976904139283597	3.1969180138839790	3.2005830260950896
METR-LA	NO_TID	2.9501271768824378	2.9491135897168523	2.9559217767245610
METR-LA	NO_DIW	2.9130885290030570	2.9094255444928240	2.9202253297254233

UTC. All formal ablations followed both submissions and the prelaunch check. All 18 formal ablations completed with zero test-open events. These times establish pre-run calendar-submission order. A later receipt upgrade anchored the P170 contract in Bitcoin block 968554, but the block time followed the first P170 result and therefore does not establish a pre-result freeze.

M.5 Eighteen seeds for the PEMS08 dropout effect

This precision and calibration study is outside the 48 registered hypotheses and asks how precisely the PEMS08 block-dropout effect is estimated and how often the frozen three-seed gate detects it under observed seed variation. The validation-only comparison holds the accepted 18-seed $p = 0.10$ control against an 18-seed $p = 0.35$ treatment: six previously accepted treatment seeds and 12 new runs on seeds 3413–3424. A zero-impact replay must first match the accepted treatment value for seed 3410 bit-exactly. Otherwise no effect label is issued. The prespecified summaries are a Welch 95% interval for the 18-versus-18 mean difference, the fraction of all $\binom{18}{3}^2 = 665{,}856$ control–treatment triplet pairs that cross the unchanged gate, and the count across six disjoint matched three-seed blocks. The triplet-pair fraction is descriptive because the pairs reuse seeds. The 18-seed estimate pools accepted cohorts from multiple contracts, as Table 33 records. Although the seed identifiers are paired, the Welch interval treats the two arms as independent. The original and fresh-seed three-run comparisons remain alongside the larger-cohort estimate, and neither the frozen P141 nor P160 label changes. P171 was frozen before its 12 new runs but designed after the three-seed results. The $p = 0.10$ arm reuses seeds 3407–3409 from the P156 old control, 3410–3412 from the P156 A/A new-seed arm, and 3413–3424 from P165. The $p = 0.35$ arm reuses P141 seeds 3407–3409 and P160 seeds 3410–3412. Only seeds 3413–3424 were run for P171. Table 33 reports every validation MAE without truncation.

Table 33: **PEMS08 validation MAE for the 18 control and 18 treatment seeds.** The final column identifies the treatment provenance; all control values were accepted under P156/P165.

Seed	$p = 0.10$ control	$p = 0.35$ treatment	Treatment source
3407	13.405660828854629	13.27424306291375	P141
3408	13.367085227396428	13.29281666821284	P141
3409	13.314486114179191	13.257271108359893	P141
3410	13.420321409350732	13.294346097727345	P160

Seed	$p = 0.10$ control	$p = 0.35$ treatment	Treatment source (continued)
3411	13.429759778648904	13.282060159617847	P160
3412	13.41968499063074	13.302069357966172	P160
3413	13.35753245509266	13.220720535436573	P171
3414	13.345082782014849	13.260441622988985	P171
3415	13.349942838364916	13.262369530305634	P171
3416	13.43313488398634	13.287989641397193	P171
3417	13.379057311260349	13.304615570193594	P171
3418	13.352361699246563	13.234638207081643	P171
3419	13.365895704811956	13.238895498262531	P171
3420	13.36111827863275	13.275505994093477	P171
3421	13.38302860359414	13.22729807557504	P171
3422	13.359787078326121	13.288282365754313	P171
3423	13.368754531919063	13.279525881703965	P171
3424	13.365548049678361	13.27559369439754	P171

The 18-seed control mean is 13.376569031443816 MAE, and the treatment mean is 13.269926837332685 MAE. The difference $d = \overline{C} - \overline{T} = 0.10664219411113152$ MAE has a Welch 95% interval of $[0.08674740615854956, 0.12653698206371347]$ (32.146838046342275 degrees of freedom). The frozen label is `EFFECT_CI_ABOVE_ZERO`.

The frozen three-seed rule triggered in all 665,856 of 665,856 control–treatment triplet pairs. Table 34 places this real-effect detection rate alongside the accepted same-configuration A/A rate and simulated one-sigma-shift rate.

Table 34: Frozen three-seed gate rates under the observed dropout effect and two accepted calibration references. The triplet pairs within the P171 column reuse seeds.

P171 real effect	P165 real A/A	Simulated one-sigma shift
100.0%	10.4%	43.9%

All six disjoint matched-seed blocks triggered ($k/6 = 6/6$). The triplet pairs overlap and therefore do not provide independent replication. This estimate concerns one effect on one dataset. It does not revise the frozen P141 or P160 decisions.

Table 35 compares the original P141 three-seed result and its P160 fresh-seed replication.

Table 35: **Three-seed results preceding the 18-seed precision study.** Gain is control minus treatment validation MAE.

Block	Gain (MAE)	Gain / $s_{\max}$
P141 original	0.08763377698125474	1.9147913566808012
P160 fresh seeds	0.13043018777300297	12.925489088670172

The contract was submitted to public OpenTimestamps calendars at 12:53 UTC. The adjudicator followed at 13:01:18 UTC. The zero-impact replay then returned 13.294346097727345 MAE for seed 3410, bit-exactly matching the accepted P160 value, before the first new-seed launch at 14:31:31.938349 UTC. All 12 new runs were validation-only and recorded zero test-open events. These calendar-submission timestamps establish pre-run order. A later receipt upgrade anchored the P171 contract in Bitcoin block 968554, whose block time preceded the first P171 result.

M.6 Rank-accounting sensitivity and paired comparisons

We reproduce the complete-ledger mean ranks of 5.528 for Orion and 2.583 for PGCRN by ranking each of the 36 dataset–horizon–metric cells and then averaging the cell ranks. Table 36 applies the same within-cell rule to the stated accounting choices. Across choices (a)–(f), Orion's mean rank ranges from 4.861 to 5.708, and its position is fifth or sixth; the dataset-specific positions are second on PEMS08, eighth on METR-LA, and sixth on PEMS-BAY.

Table 36: **Mean-rank sensitivity to stated accounting choices.** Each cell rank is one plus the number of strictly smaller values, so tied values share a rank. The top-five column gives model names followed by mean ranks; row (c) averages MetaDG over its 24 non-METR-LA cells and every other model over all 36 cells. †AutoSTF-NoNAS omits the official NAS architecture weights.

Accounting choice	Cells	Orion rank / position	Top five by mean rank
(a) Original, 12 models	36	5.528 / 5th	PGCRN 2.583; HimNet 3.889; D2STGNN 4.222; STAEformer 4.639; Orion 5.528
(b) Remove AutoSTF-NoNAS†	36	4.861 / 5th	PGCRN 2.583; HimNet 3.750; D2STGNN 3.806; STAEformer 4.194; Orion 4.861
(c) Omit MetaDG on METR-LA	36	5.528 / 6th	MetaDG 2.583; PGCRN 2.583; HimNet 3.889; D2STGNN 4.222; STAEformer 4.639
(d) Remove MetaDG	36	4.861 / 5th	PGCRN 2.250; HimNet 3.528; D2STGNN 3.833; STAEformer 4.028; Orion 4.861
(e) MAE and RMSE only	24	5.708 / 6th	PGCRN 2.500; HimNet 4.083; D2STGNN 4.292; STAEformer 4.583; AutoSTF-NoNAS† 5.375
(f) (b) and (e)	24	5.042 / 5th	PGCRN 2.500; D2STGNN 3.875; HimNet 3.875; STAEformer 4.083; Orion 5.042
(g1) PEMS08 only	12	2.000 / 2nd	MetaDG 1.000; Orion 2.000; STAEformer 4.000; HimNet 4.250; PGCRN 4.333
(g2) METR-LA only	12	8.250 / 8th	D2STGNN 1.250; PGCRN 2.417; STAEformer 3.500; ST-SSDL 3.583; HimNet 4.250
(g3) PEMS-BAY only	12	6.333 / 6th	PGCRN 1.000; D2STGNN 2.833; HimNet 3.167; MetaDG 4.167; AutoSTF-NoNAS† 4.333

Among the 15 rows of Table 27 with both a computable paired interval and a recorded frozen support-trigger field, "threshold crossed" and "interval excludes zero" agree in 11 rows and differ in four. The four differences are PEMS08 width 96, PEMS08 width 128, PEMS08 graph learning rate 0.10, and PEMS08 block dropout 0.35; the other three rows of Table 27 lack a three-pair interval and are excluded from this comparison.

N Claim-to-Evidence Map

This map follows the abstract and the closely linked headline claims in Sections 3–7. A hypothesis is a registered question, an arm is one contracted comparison, and a display cell is one metric–dataset–horizon entry. These units are not interchangeable. "Partial" chronology means that local hashes and event records do not provide an external pre-result witness for every earlier contract. A validation verdict does not turn a later test replay into an untouched test. The P165 calibration appears in the intervention registry. The additional audit cases listed after the divider are outside the existing 48 registered Orion hypotheses. The P171 triplet-pair rate reuses seeds and is descriptive; prior frozen labels stand. Here "rerun" means an independent byte-level recomputation of the adjudication artifact, not an independent scientific replication. Table 37 maps each claim to its evidence unit.

Table 37: **Abstract claims and their evidence units at the current accepted snapshot.** Where combines Main text with Figure or table; Evidence record combines Contract or record ID with Unit; Status combines Time chain, Validation verdict, Test exposure, and Rerun. "NR" means no accepted result is reported for that field; "n/a" means the field does not apply. P171's overlapping triplet pairs give a descriptive rate.

Claim	Where	Evidence record	Status
Comparator omission changes a lead with fixed forecasts	*text:* §6.3; §7.1; *display:* Table 1; Fig. 11	*record:* closure record; P151; *unit:* display cell	*time:* after complete roster; partial; *val.:* n/a; eligibility audit; *test:* test ledger disclosed; *rerun:* recomputed

Claim (continued)	Where	Evidence record	Status
Across twelve leave-one-comparator-out analyses, six of twelve models gain at least one highlight; omitting MetaDG specifically moves Orion from 0/36 to 12/36	*text:* Introduction; §6.3; §7.1; *display:* Fig. 1; Table 1; Fig. 11	*record:* closure record; P151; *unit:* model over display cells	*time:* after roster closure; *val.:* n/a; audit; *test:* test values reused descriptively; *rerun:* recomputed
Pre-result contract fixes each proposed change and its rule	*text:* §3; *display:* Fig. 2; Table 4	*record:* registry; chronology scope; *unit:* hypothesis	*time:* 8 pre-result calendar submissions; 40 without a pre-result external anchor; *val.:* per-contract ruling; *test:* per-contract ledger; *rerun:* NR as a set
Contracts bind each change and complete scripted rulings are replayable; the ledger documents 71 known logical events but is not an exhaustive access history	*text:* §3; §7.1; *display:* Fig. 2; Tables 3 and 4; App. F	*record:* registry; closure record; exposure ledger; *unit:* hypothesis; arm; exposure event	*time:* per-contract; partial; *val.:* fail-closed scripted ruling; documented review exception; *test:* 71 events; 15 dates and 50 later-use fields unresolved; *rerun:* per scripted record
Study-level preregistration can specify individual ablations; our contribution operationalizes each change with a matched control, executable decision rule, and auditable evidence bundle	*text:* Related Work; App. B; *display:* Table 2	*record:* boundary table; *unit:* evaluation practice	*time:* published reports; current accepted snapshot; *val.:* reported-scope comparison; *test:* n/a; *rerun:* n/a
Zero-impact, seed, comparator, and stop checks precede a label	*text:* §3; *display:* Table 3	*record:* P131; P140; closure record; *unit:* arm	*time:* per-contract; partial; *val.:* stop or complete record; *test:* no test on documented stops; *rerun:* control replays
One post-selection test replay per closed recipe; earlier access logged	*text:* §3; §7.1; *display:* Table 1; App. F	*record:* exposure ledger; closure record; *unit:* recipe; exposure event	*time:* prior access retained; *val.:* recipe fixed on validation; *test:* one replay; prior access disclosed; *rerun:* n/a
Normal-model null trigger 10.6%; one-SD detection 44%	*text:* §5; *display:* Fig. 4	*record:* P154 acceptance; protocol macros; *unit:* arm comparison	*time:* simulation; n/a; *val.:* descriptive calibration; *test:* none; *rerun:* independent simulation check
Under the null, the prespecified treatment has the lower three-seed mean in 50% of comparisons	*text:* §5; *display:* Fig. 4	*record:* P163 acceptance; protocol macros; *unit:* arm comparison	*time:* simulation; n/a; *val.:* descriptive reference; *test:* none; *rerun:* independent simulation check
Real-seed null partitions trigger in 10.4%	*text:* §5; *display:* Fig. 4; Table 8	*record:* P165 acceptance; *unit:* dependent arm split	*time:* pre-run sidecar and OpenTimestamps (OTS); *val.:* gate-passed calibration verdict (observed null-split rate within the prespecified band); not an intervention-efficacy finding; *test:* validation only; *rerun:* adjudicator re-run
The three-seed rule is a decision convention, not a significance test	*text:* §5; *display:* Fig. 4; Table 7	*record:* P154 acceptance; protocol macros; *unit:* arm comparison	*time:* simulation; n/a; *val.:* descriptive calibration; *test:* none; *rerun:* independent simulation check
One observational probe and three diagnostic-led intervention tests do not meet their frozen action criteria under the evaluated settings	*text:* §6.1; *display:* Fig. 5; Tables 16–19	*record:* powered diagnosis; P131; P126 early; P140; *unit:* hypothesis or arm	*time:* partial; probe threshold caveat; *val.:* inconclusive or below the frozen gain criterion; *test:* validation evidence; *rerun:* NR as a set
Normalization and weight decay remain below their frozen gain criteria	*text:* §6.1; *display:* Fig. 5; Table 4	*record:* registry; closure record; *unit:* arm	*time:* partial; *val.:* below the frozen gain criterion; *test:* validation evidence; *rerun:* NR

Claim (continued)	Where	Evidence record	Status
Raising dropout from 0.10 to 0.35 helps PEMS08, not PEMS-BAY or METR-LA	*text:* §6.2; *display:* Fig. 5; Fig. 7	*record:* P141; P160 acceptance; registry; *unit:* arm by dataset	*time:* contract; P160 pre-run OTS; *val.:* PEMS08 gain; other scopes flat or worse; *test:* validation only for fresh cohort; *rerun:* P160 fresh cohort
All three recipe changes tested on fresh seeds pass again; other recipe components were not tested on fresh seeds	*text:* §6.2; *display:* Fig. 5	*record:* P160; P168 acceptances; *unit:* three arms	*time:* contracts precede fresh runs; *val.:* same frozen gate passed; *test:* fresh cohorts validation only; *rerun:* fresh cohorts
Selected PEMS08 recipe: 8 of 36 cross-dataset validation cells, all on PEMS08, versus 0 of 36 test leads	*text:* §7.1; *display:* Fig. 2(e); Table 1	*record:* closure record; ledger macros; *unit:* display cell	*time:* recipe fixed before replay; earlier access logged; *val.:* selected on validation; *test:* post-selection replay; *rerun:* n/a
Orion is second only to MetaDG in every PEMS08 test cell and ranks fifth of twelve overall	*text:* §6.3; §7.1; *display:* Tables 1 and 22	*record:* closure record; full-horizon ledger; *unit:* display cell; model rank	*time:* after complete roster; earlier access logged; *val.:* n/a; complete-ledger ranking; *test:* test ledger disclosed; *rerun:* recomputed
120/124 same-metric reports lie within the descriptive band; protocol equivalence is not implied	*text:* §7.1; *display:* Table 24; Fig. 12	*record:* closure record; P153/P166; *unit:* reported metric cell	*time:* source and reproduction logged; *val.:* n/a; *test:* reproduced test metrics; *rerun:* reproduction ledger
48 registered hypotheses and their records are supplied	*text:* §3; *display:* Table 4	*record:* registry; closure record; *unit:* hypothesis	*time:* partial across historical set; *val.:* includes stops and withdrawals; *test:* exposure ledger supplied; *rerun:* per record
Eight of the 48 contracts have Bitcoin block-header times before their first results or no-run bounds; five precede them by more than the two-hour tolerance	*text:* §3; *display:* Fig. 1; Table 25	*record:* chronology scope; OTSCONFIRM acceptance; *unit:* contract	*time:* 8 header-before, including 5 beyond tolerance; 22 after; 4 without a strict time comparison; 14 without mapped confirmation; *val.:* n/a; chronology audit; *test:* n/a; *rerun:* independent recount and block check
Contracts, adjudicators, outcomes, and code are in the supplementary material	*text:* Reproducibility statement; App. D; *display:* Table 4	*record:* supplementary README and REPRODUCE files; registry; *unit:* artifact set	*time:* current accepted snapshot; *val.:* acceptance records supplied; *test:* exposure ledger supplied; *rerun:* per record
Additional cases and audits outside the registered Orion-hypothesis denominator			
Two known-harm reversals recovered on PEMS08	*text:* §5; *display:* Table 29	*record:* P169 acceptance; *unit:* arm contrast	*time:* pre-run calendar; block before first result; *val.:* both RECOVERED; gains 0.927/0.175 MAE; *test:* no test opening; *rerun:* independent identical rerun
STID implementation components: four gate-passed, one flat, one reversed	*text:* §3; §7.1; *display:* Table 31	*record:* P170 acceptance; *unit:* arm by dataset; second model	*time:* pre-run calendar; block after first result; *val.:* six complete rulings: 4/1/1; *test:* validation only; no test; *rerun:* independent identical rerun
PEMS08 dropout: 18-seed gain with CI above zero	*text:* §6.2; *display:* App. M.5; Table 34	*record:* P171 acceptance; *unit:* 18 seeds/arm; triplet pair	*time:* pre-run calendar; block before first result; *val.:* gain 0.107 MAE; CI [0.0867, 0.1265]; 100% pairs; 6/6 blocks; *test:* new runs validation only; zero test opens; *rerun:* independent byte-identical rerun
Bitcoin anchoring separates pre- and post-result receipts	*text:* §3; *display:* Table 25	*record:* OTS CONFIRM acceptance; *unit:* registered contract	*time:* 8 header-before, 5 beyond the two-hour tolerance; 22 after; 4 unknown; 14 none; *val.:* n/a; receipt audit; *test:* n/a; *rerun:* independent recount and block check

Claim (continued)	Where	Evidence record	Status
Post hoc descriptive BH flags 8 at $q = 0.10$ (6 at $q = 0.05$), and Holm flags 4 at $\alpha = 0.05$; frozen labels stay unchanged	*text:* §5; *display:* Tables 26, 27	*record:* FDRPAIR acceptance; *unit:* decision family; pair	*time:* after frozen verdicts; *val.:* descriptive flags only; labels unchanged; *test:* no test artifacts opened; *rerun:* independent BH/Holm recomputed

The displayed numbers use the protocol, closure-ledger, and chronology records in the supplementary material, plus the accepted records named in the record column. P151 and P153/P166 denote the accepted opponent-sensitivity and reproduction records used by the closure ledger. The four evaluation sources are the powered diagnosis and the P131, P126 early-activation, and P140 acceptances.

The anonymous supplementary material links each P169–P171, OTSCONFIRM, and FDRPAIR record to its accepted label, numerical summary, provenance, and independent verification status; these additional cases remain outside the 48-hypothesis Orion denominator. The P169, P170, and P171 contracts were submitted to public calendars before their runs; a later upgrade anchored all three in Bitcoin block 968554, whose time precedes the first P169 and P171 results but follows the first P170 result (Appendix M.1). P171 was frozen before its 12 new runs but designed after the three-seed results. The new runs were validation-only with zero test-open events. Its existing P141/P160 frozen labels remain unchanged.

Figure credits. The pictograms in Figures 2, 3, and 9 are from Flaticon.com.