MIRAGEBENCH: WHEN DOES CAUSAL DISCOVERY SURVIVE LATENT CONFOUNDING?

Anonymous authors
Paper under double-blind review

ABSTRACT

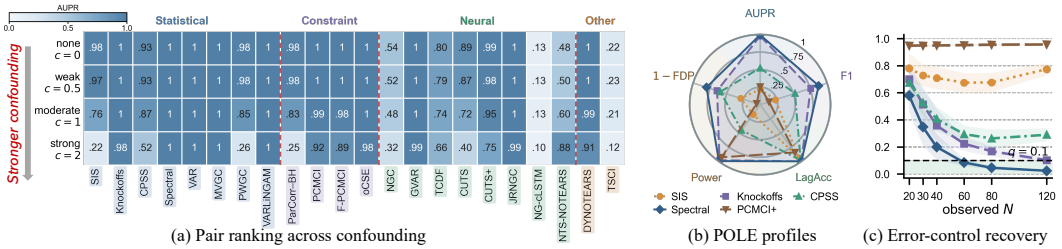
Latent confounding occurs when an unobserved common cause influences multiple observed variables, inducing dependence that can be mistaken for a direct causal relation. Yet the reliability of causal discovery under increasing confounding remains difficult to assess. In this work, we study this problem in observational multivariate time series, where direct lagged relations provide an explicit structural target. We introduce MIRAGEBENCH, which preserves lagged-edge ground truth across confounding strengths and also varies the number of observed variables, record length, and the number of latent drivers. To compare heterogeneous method outputs, we introduce POLE, an output-aware protocol that separately evaluates **P**air ranking, **O**perating points, **L**ag localization, and **E**rror control. Controlled experiments reveal three findings. First, under latent confounding, methods exhibit heterogeneous responses across evaluation layers, and near-perfect pair ranking can coexist with substantial false discoveries. Second, under strong confounding, several selectors show lower false-discovery proportions at wider observed cross-sections, sometimes at a cost in edge power, whereas longer records at a fixed cross-section do not yield comparable error reduction. Third, within the evaluated linear-Gaussian generator, recovery depends on the selector, the number of latent drivers under the benchmark parameterization, and observational resources. Together, these results show that *reliability under latent confounding depends on what is evaluated, which selector is used, and what observational resources are available*. Our code and benchmark are available at `https://anonymous.4open.science/r/miragebench-paper`.

1 INTRODUCTION

Causal discovery aims to infer causal structure from observational data (Assaad et al., 2022) and is valuable in domains such as climate science (Runge et al., 2019a), neuroscience (Smith et al., 2011), and epidemiology (Deyle et al., 2016), where controlled interventions are often costly, constrained, or infeasible. In these complex systems, however, the relevant processes are rarely fully observed: important common drivers may remain unmeasured. This gives rise to latent confounding, where an unobserved common cause influences multiple observed variables, creating associations that may be mistaken for direct causal relations (Spirtes et al., 2000). This raises our central question: *How reliable is causal discovery under latent confounding?* Addressing this question empirically calls for two complementary benchmark requirements: **(1) controlled variation in latent confounding**, so that methods can be compared against the same underlying causal target as confounding strengthens; and **(2) evaluation that respects heterogeneous structural outputs**, since methods may return pair scores, selected graphs, or lag-resolved relations that support different structural claims.

We study this question in an *observational multivariate time series* setting, where direct lagged relations provide an explicit structural target. We vary two *observational resources*: the number of observed variables N, also called the observed cross-section, and the record length T. Existing time series benchmarks expose known graphs under configurable mechanisms (Lawrence et al., 2021; Ferdous et al., 2025; Herdeanu et al., 2025) or examine graph recovery under graded or scenario-based violations of modelling assumptions (Stein et al., 2026; Yi et al., 2026). However, these efforts do not jointly provide the two properties above: holding the direct-edge target fixed while varying latent confounding, and evaluating the distinct structural claims supported by heterogeneous outputs. This distinction matters because a common pair-level target can support shared comparisons,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Pair ranking across confounding (b) POLE profiles (c) Error-control recovery

Figure 1: Overview of MIRAGEBENCH under latent confounding. (a) Pair AUPR by confounding strength and method family. (b) Five selectors across POLE under strong confounding: pair AUPR (P), reported-set pair F1 (O), conditional lag accuracy (L), and descriptive edge power and $1 - \mathrm{FDP}$ (E). (c) Mean realized FDP as the number of observed variables increases under strong confounding.

whereas lag placement and error control require information that a pair projection does not retain. Once these requirements are met, a further empirical question is whether additional observational resources can mitigate the effects of confounding.

To address the first requirement, we introduce MIRAGEBENCH, a benchmark for causal discovery under controlled latent confounding, where hidden common causes can create a "mirage" of direct causal relations absent from the ground truth. Specifically, the generator provides exact ground truth for direct lagged relations among observed variables. We hold this structure fixed while varying confounding strength, so that methods are evaluated against the same causal target across confounding levels. Figure 1a reveals heterogeneous responses: pair-ranking performance is nearly preserved for some methods, declines sharply for others, and improves in some cases.

To address the second requirement, we introduce POLE, an output-aware four-layer evaluation framework comprising **P**air ranking, **O**perating points, **L**ag localization, and **E**rror control. Each method is evaluated at the layers supported by its returned output. Figure 1b illustrates why these distinctions matter: under strong confounding, *good performance in one layer need not extend to the others*. For example, Spectral deconfounding (Ćevid et al., 2020) retains near-perfect pair ranking (AUPR 0.996), yet its selected set has $1 - \mathrm{FDP} \approx 0.80$, meaning that about 20% of the reported edges are false on average, which is twice the benchmark target $q = 0.1$.

Beyond comparing method performance, MIRAGEBENCH also enables controlled experiments for us to study how observational resources affect discovery under latent confounding and when error control can recover. Figure 1c illustrates the cross-section response using FDP (False Discovery Proportion, the fraction of selected edges that are false) as the number of observed variables increases. Mean FDP decreases with the observed cross-section for several selectors, and Spectral deconfounding meets our empirical recovery criterion in this fixed-T comparison (Section 4.2). Longer records, however, show no comparable FDP reduction at a fixed cross-section in our experiments. For spectral deconfounding on the crossed grid, the number of observed variables required for recovery increases with record length, and at the longest record none of the evaluated selectors recovers within the tested range (Section 4.2). To summarize, our contributions are threefold:

- **A benchmark with controlled confounding.** We introduce MIRAGEBENCH, which provides exact direct lagged-edge ground truth while varying confounding strength, observational resources, and the number of latent drivers.
- **Output-aware, multi-layer evaluation.** We introduce POLE, which maps compatible outputs to a shared pair-level view while separately evaluating pair ranking, operating points, lag localization, and error control wherever the returned output supports them.
- **Benchmark-enabled scientific findings.** Controlled experiments reveal: (i) under confounding, methods exhibit heterogeneous responses across layers, with near-perfect pair ranking yet false discoveries; (ii) under strong confounding, several selectors have lower FDP at wider cross-sections, sometimes at a cost in edge power, whereas longer records at a fixed cross-section offer no comparable FDP reduction; and (iii) recovery depends on selector, driver count under the benchmark parameterization, and observational resources.

Table 1: Design dimensions in related benchmarks. See Appendix E.1 for criteria and qualifications.

Benchmark	Ground truth	Latent confounding	Graded confounding	N and T variation	Multi-resolution evaluation	Error-target assessment
PCMCI tasks (Runge et al., 2019b)	✓					
CauseMe/C4C (Runge et al., 2020)	✓	✓		✓		
Lawrence et al. (2021)	✓	✓	✓	✓		
SAVAR (Tibau et al., 2022)	✓			✓		
CausalTime (Cheng et al., 2024b)	✓	✓				
Industrial (Mogensen et al., 2024)						
TimeGraph (Ferdous et al., 2025)	✓	✓				
CausalDynamics (Herdeanu et al., 2025)	✓	✓				
CausalRivers (Stein et al., 2025)		✓		✓		
CausalProfiler (Panayiotou et al., 2025)						
TCD-Arena (Stein et al., 2026)	✓	✓	✓	✓	✓	
CausalCompass (Yi et al., 2026)	✓	✓	✓	✓		
MIRAGEBENCH (ours)	✓	✓	✓	✓	✓	✓

2 RELATED WORK

Causal discovery under latent confounding. Causal discovery methods differ both in their assumptions about latent confounding and in the structural objects they return. PCMCI and PCMCI+ assume causal sufficiency and return lagged links (Runge et al., 2019b; Runge, 2020), whereas LPCMCI, SVAR-FCI, and TS-ICD accommodate latent confounders and return partial ancestral or ancestral relations (Gerhardus & Runge, 2020; Malinsky & Spirtes, 2018; Yehezkel Rohekar et al., 2023). Other approaches rely on additional structure: tensor-rank conditions can identify causal structure in discrete time series with latent confounders (Chen et al., 2027), while joint analysis across multiple datasets can recover structure in the presence of partially observed contexts (Günther et al., 2023). Factor-augmented VARs and low-rank structured-sparse VARs also model latent structure in high-dimensional dynamics (Bernanke et al., 2005; Basu et al., 2019), but target factor-augmented structural analysis or sparse interaction recovery rather than the benchmark's formal error-control target over direct lagged-edge sets. Score-based and neural time series methods instead expose lagged coefficients, pair-level scores, or learned structures under their own modelling assumptions (Pamfil et al., 2020; Nauta et al., 2019; Tank et al., 2022). These outputs are not interchangeable and support different empirical claims, motivating evaluation that respects the information each method actually returns.

Time series causal discovery benchmarks. Benchmark designs trade experimental control against realism. Controlled synthetic suites provide known graphs and configurable mechanisms (Lawrence et al., 2021; Ferdous et al., 2025; Herdeanu et al., 2025), while assumption-violation benchmarks systematically vary departures from modelling assumptions (Stein et al., 2026; Yi et al., 2026). We also include adjacent works for context: a stochastic climate model used to benchmark teleconnection recovery (Tibau et al., 2022), a configurable causal-data generator without a fixed reported evaluation (Panayiotou et al., 2025), and synthetic tasks accompanying a method paper (Runge et al., 2019b). Realism-oriented benchmarks use synthetic data designed to approximate observed dynamics, paired with derived graphs (Cheng et al., 2024b), or real records with constructed or expert-specified references (Stein et al., 2025; Mogensen et al., 2024). Mixed suites combine synthetic, hybrid, and real series (Runge et al., 2020). Table 1 compares these works across six design dimensions (Appendix E.1). TCD-Arena is the closest existing benchmark in scope. MIRAGEBENCH additionally combines controlled latent confounding with crossed variation in observed-variable count and record length under fixed direct-edge ground truth, and assesses an error-control target in prespecified method–setting families rather than claiming a theoretical FDR guarantee.

3 MIRAGEBENCH: BENCHMARK DESIGN

To study causal discovery under controlled latent confounding while comparing heterogeneous method outputs fairly, MIRAGEBENCH combines a generator that holds the direct-edge target fixed as the latent term varies with the output-aware POLE evaluation protocol (Figure 2). We first define the task and then present the generator and evaluation protocol in the following subsections, respectively. Algorithms 1–3 give the complete procedures.

Task and targets. We here study causal discovery from observational multivariate time series as a structured setting for evaluating the effects of latent confounding. The observed record $\mathbf{X} =$

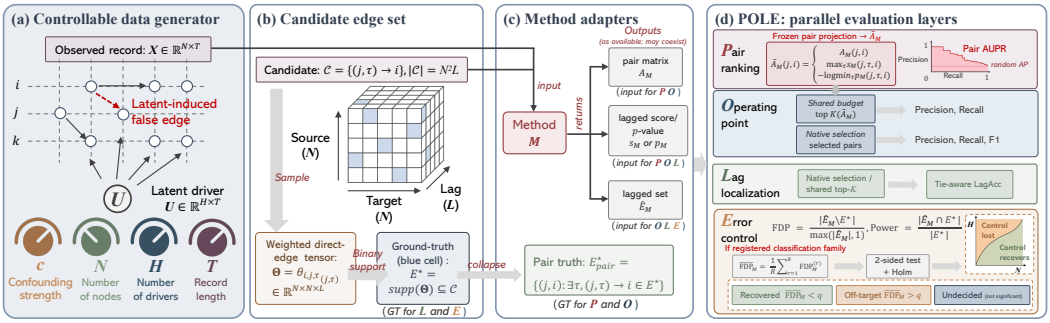


Figure 2: MIRAGEBENCH workflow. GT: ground truth. AP: average precision.

$\{x_{i,t}\} \in \mathbb{R}^{N\times T}$ contains N observed variables measured over T time steps and may be influenced by H unobserved latent drivers $\mathbf{U} = \{u_{h,t}\} \in \mathbb{R}^{H\times T}$. A candidate edge $(j,\tau) \to i$ posits a direct effect of variable j on variable i after τ time steps, where L is the maximum candidate lag, giving the lag-resolved candidate universe

$$\mathcal{C} \;=\; \{(j,\tau) \to i : 1 \le i,j \le N,\; 1 \le \tau \le L\}, \qquad |\mathcal{C}| = N^2 L. \tag{1}$$

Let $E^\star \subseteq \mathcal{C}$ denote the benchmark target, the set of ground-truth observed-to-observed direct lagged edges. For pair-level evaluation, the cross-node pair universe and its collapsed truth are

$$\mathcal{U}_{\text{pair}} = \{(j,i) : 1 \le i,j \le N,\; j \ne i\}, \qquad E^\star_{\text{pair}} = \{(j,i) \in \mathcal{U}_{\text{pair}} : \exists\tau,\; (j,\tau) \to i \in E^\star\}. \tag{2}$$

Self-lagged edges remain in the lag-resolved universe, whereas the common pair-level comparison is restricted to cross-node pairs and self-pairs are evaluated separately (Appendix Table 5). Our target is restricted to direct lagged edges; instantaneous effects, interventional distributions, and ancestral relations are outside the scope.

3.1 CONTROLLABLE GENERATOR

Direct-edge structure. A sparse coefficient tensor $\mathbf{\Theta} = \{\theta_{i,j,\tau}\}$ assigns lagged direct effects over the candidate universe $\mathcal{C}$ of Eq. 1. The generator instantiates the target $E^\star$ through the support of $\mathbf{\Theta}$, whose entry $\theta_{i,j,\tau}$ is the coefficient on $x_{j,t-\tau}$ in the equation generating $x_{i,t}$ below:

$$E^\star = \{(j,\tau) \to i \in \mathcal{C} : \theta_{i,j,\tau} \ne 0\}. \tag{3}$$

Latent confounding term. The generator samples H independent latent drivers, each following a stable first-order autoregressive process (AR(1)):

$$u_{h,t} \;=\; \rho\, u_{h,\,t-1} \;+\; \eta_{h,t}, \qquad \eta_{h,t} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad |\rho| < 1. \tag{4}$$

Each driver reaches a configured group of observed variables, and for each variable in that group the generator samples a signed loading $b_{i,h}$ and a lag $\ell_{i,h} \in \{1,\dots,L\}$, with $b_{i,h} = 0$ outside the group. Because the lag is variable-specific, one driver can appear at different offsets across variables, creating an apparent directed lagged association where no direct edge exists, while the scalar $c \ge 0$ scales the latent confounding term.

Observed dynamics. Given the direct-edge structure and the latent drivers above, each observed variable $x_{i,t}$ follows a linear-Gaussian dynamic model:

$$x_{i,t} \;=\; \sum_{\tau=1}^{L} \sum_{j=1}^{N} \theta_{i,j,\tau}\, x_{j,\,t-\tau} \;+\; c \sum_{h=1}^{H} b_{i,h}\, u_{h,\,t-\ell_{i,h}} \;+\; \varepsilon_{i,t}, \tag{5}$$

Here $\varepsilon_{i,t} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\varepsilon)$ is independent observation noise. The first term follows the structure encoded by $\mathbf{\Theta}$, while the second is the latent confounding term, which adds latent variation without changing $\mathbf{\Theta}$ or $E^\star$. Setting $c = 0$ removes it, and larger c strengthens its contribution. Appendix B.1 details edge sampling, stability control, and the generator variants.

Scope of the generator. This setting is a controlled test bed rather than a realism claim. It provides exact direct lagged-edge truth, an explicitly parameterized latent confounding term scaled by c, and independent control over observational resources and the number of latent drivers. These choices hold the observed-to-observed direct structure fixed across confounding levels before adding non-linear dynamics or domain-specific structure.

Table 2: The four POLE layers. $\mathrm{AP}_{\mathrm{rand}}$ is the random-ranking reference, and Appendix B.3 gives the full admission and scoring rules.

Layer	Required output	Scored object	Truth	Metrics
Pair ranking	pair score	full pair ranking $\widetilde{A}_M$	$E^\star_{\mathrm{pair}}$	pair AUPR; $\mathrm{AP}_{\mathrm{rand}}$
Operating points	pair score or selected set	shared budget: $\mathrm{TopK}_{K=N}(\widetilde{A}_M)$; reported set: selected pairs	$E^\star_{\mathrm{pair}}$	precision, recall; selected-set F1
Lag localization	lag information	true pairs in Q_L	$E^\star$	LagAcc; denominator
Error control	reported lagged set	$\widehat{E}_M$	$E^\star$	FDP, power

3.2 POLE EVALUATION

A method may return structural scores, selected relations, or lag-resolved information. *Pair-level evaluation* asks whether a direct relation exists at any candidate lag, whereas *lag-level evaluation* additionally asks whether the correct delay is identified. A returned object enters every layer its capabilities support, and the four layers are reported separately with no composite score (Table 2).

Pair ranking and operating points. For score-bearing outputs, a *frozen projection* fixed before evaluation constructs the pair-level score $\widetilde{A}_M$: it takes the maximum over lags for per-lag strengths, $-\log$ of the minimum for per-lag p-values, and native pair scores directly. Pairs without a returned score receive that method's worst score, so every projected ranking covers the same cross-node pair universe $\mathcal{U}_{\mathrm{pair}}$. The P layer evaluates the full projected ranking against $E^\star_{\mathrm{pair}}$. The O layer evaluates its top-$K{=}N$ pairs and, separately, the pair projection of an available reported selected set.

Lag localization. The evaluated pair set Q_L is formed from the pair projection of a reported selected set when available and otherwise from the top-$K{=}N$ projected pair ranking. Lag accuracy is computed only over the true pairs in Q_L, with lags tied for best sharing credit, and the number of such pairs is reported as the denominator. A zero denominator leaves lag accuracy missing rather than zero. Algorithm 3 gives the full tie and missing-output rules.

Error control. The E layer evaluates a reported lagged-edge set $\widehat{E}_M$ against the exact lag-resolved truth $E^\star$. We use $q \in (0, 1)$ as the benchmark reference error level, and a method's native nominal target, when present, is recorded separately.

Definition 1 (False-discovery metrics)**.** *Assume $E^\star \neq \varnothing$. For any reported edge set $\widehat{E} \subseteq \mathcal{C}$, the realized **False Discovery Proportion (FDP)** and **power** are*

$$\mathrm{FDP}(\widehat{E}) = \frac{|\widehat{E} \setminus E^\star|}{\max(|\widehat{E}|, 1)}, \qquad \mathrm{Power}(\widehat{E}) = \frac{|\widehat{E} \cap E^\star|}{|E^\star|}. \tag{6}$$

*The **False Discovery Rate (FDR)** of M is $\mathrm{FDR}_M = \mathbb{E}\Big[\mathrm{FDP}(\widehat{E}_M)\Big]$, where the expectation is over repeated benchmark instances and any method randomness at a fixed configuration. At the fixed benchmark configuration, the FDR target is met when $\mathrm{FDR}_M \leq q$.*

3.3 RECOVERY CLASSIFICATION PROTOCOL

Mean realized FDP estimates FDR_M, but a mean below q does not establish recovery. Formal classification is restricted to prespecified method–setting families of E-layer selectors that return lag-resolved direct-edge sets under the common FDP/FDR estimand and reference level q. At each setting in such a family, a multiplicity-corrected two-sided test on run-level FDP values classifies the selector as *recovered* when the evidence places FDR_M below q, *off-target* when it places it above q, and *undecided* otherwise. These empirical verdicts do not imply selector-level control guarantees.

Definition 2 (Empirical recovery threshold)**.** *Fix T, c, q, and a prespecified classification family for selector M. At fixed H, if the recovered set is nonempty and monotone over evaluated N, the empirical recovery threshold is $N^\star_M(H; T, c, q) = \min\{N : (N, H)$ is classified recovered$\}$, with the minimum over evaluated N. The threshold is family-dependent through multiplicity correction.*

Otherwise, no scalar threshold is reported (Appendix A.1). $N^\star_M$ concerns error control only; we examine selected-set size and edge power separately when assessing whether recovery is vacuous (Section 4.3).

4 EXPERIMENTS

We organize the experiments around three Research Questions (RQs).

- **RQ1.** How does causal discovery respond to increasing latent confounding across evaluation layers?
- **RQ2.** When discovery degrades under confounding, which additional observations help?
- **RQ3.** What determines whether and where error control can recover?

Experimental settings. The headline experiments use pervasive latent confounding, with hidden drivers loading on all observed variables and maximum lag $L = 2$ unless stated otherwise. Each RQ uses a controlled slice rather than a full factorial design, with the remaining parameters fixed within each comparison (Appendix Table 10). Unless stated otherwise, each setting is evaluated over 30 runs, and every metric is computed within a run before its mean and variability are summarized across runs. Appendix C gives the per-result settings and implementation details.

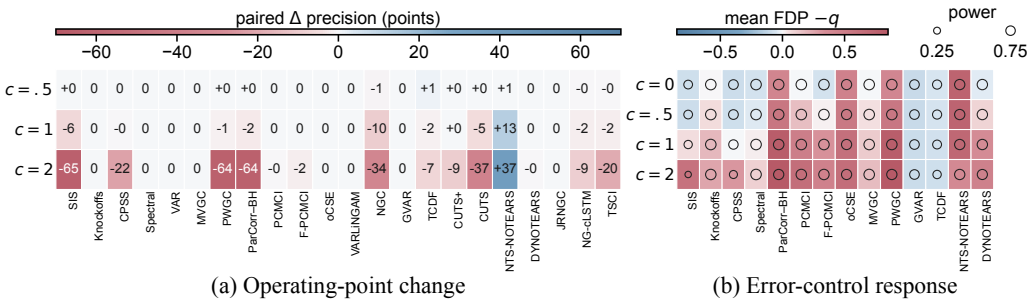
Compared methods. We quantitatively compare 22 methods across four families, detailed in Appendix Table 6. The Statistical group comprises SIS (Fan & Lv, 2008), Knockoffs (Barber & Candès, 2015; Candès et al., 2018), CPSS (Shah & Samworth, 2013), Spectral deconfounding (spectral transform followed by debiased Lasso and pooled BH at q_M=0.1) (Ćevid et al., 2020; Guo et al., 2022), VAR and pairwise Granger causality (PWGC) (Granger, 1969), multivariate Granger causality (MVGC) (Barnett & Seth, 2014), and VARLiNGAM (Hyvärinen et al., 2010). The Constraint group comprises PCMCI (Runge et al., 2019b), F-PCMCI (Castri et al., 2023), oCSE (Sun et al., 2015), and a partial-correlation–BH (ParCorr–BH) comparator implemented in this benchmark, which tests each candidate edge by partial correlation conditioning on the target's own lags and applies Benjamini–Hochberg at $\alpha = 0.1$. The Neural group comprises Neural Granger in its cMLP and cLSTM variants (Tank et al., 2022) (NGC and NG-cLSTM in the figures), GVAR (Marcinkevičs & Vogt, 2021), TCDF (Nauta et al., 2019), CUTS (Cheng et al., 2023), CUTS+ (Cheng et al., 2024a), JRNGC (Zhou et al., 2024), and NTS-NOTEARS (Sun et al., 2021), and the Other group comprises DYNOTEARS (Pamfil et al., 2020) and TSCI (Butler et al., 2024).

Evaluation protocol. We use the POLE metrics of Section 3.2, with the shared operating-point budget fixed at K=N. Formal recovery verdicts follow Section 3.3 and are restricted to the pre-specified method–setting families used in Sections 4.2 and 4.3, and all other results are descriptive. Appendix A.1 gives the full family definitions and multiplicity corrections.

4.1 MULTI-LAYER RESPONSES TO CONFOUNDING

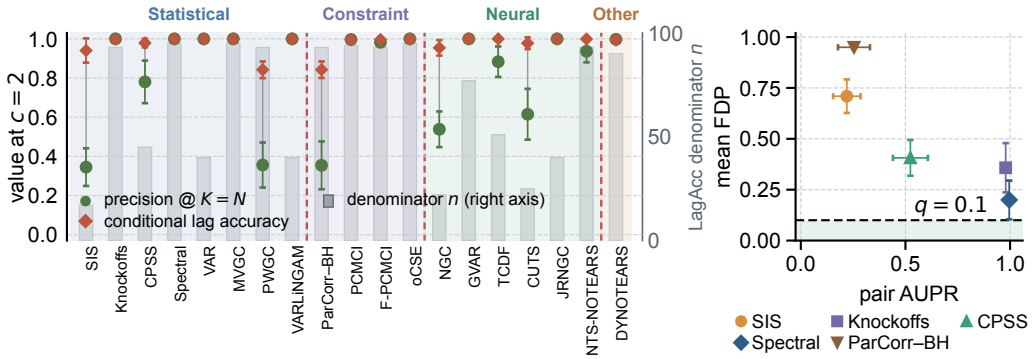
To address **RQ1**, we vary confounding strength c over four levels at the headline setting N=40, H=1, T=600, and L=2, and evaluate each output at the POLE layers it admits. Within each run, the layout, $\mathbf{\Theta}$, and $E^\star$ are shared across strengths, so the comparison is structure-matched, and the c=0 reference uses an independent noise realization while the positive strengths share one. We first report the response within each layer and then ask whether these responses agree.

Pair ranking. Pair-ranking responses are highly heterogeneous across confounding strengths (Figure 1a). Several methods retain near-ceiling AUPR even at c=2, whereas SIS, ParCorr–BH, and



(a) Operating-point change (b) Error-control response

Figure 3: Within-layer responses to confounding. (a) Change in pair precision at the shared K=N operating point relative to matched c=0. (b) Realized FDP − q for reported lagged-edge sets, with circle area indicating edge power; panel (b) is descriptive.



(a) O versus L at c=2

(b) Pair ranking and FDP

Figure 4: Cross-layer contrasts under strong confounding, c=2. (a) Shared-budget pair precision and conditional lag accuracy, with gray bars showing the mean LagAcc denominator. (b) Pair AUPR versus mean realized FDP. Error bars denote one standard deviation (SD).

several pairwise or marginal alternatives deteriorate sharply, and NTS-NOTEARS improves on this grid. The observed ordering aligns more closely with what each score conditions on than with broad method family: within the Statistical group, marginal or pairwise scores such as SIS and PWGC deteriorate, whereas Knockoffs and spectral deconfounding remain near the ceiling. We treat this as an empirical association rather than a mechanism claim.

Operating points. Precision at the shared budget $K{=}N$ changes little for several methods as confounding strengthens (Figure 3a). By contrast, SIS, ParCorr–BH, and PWGC lose more than sixty percentage points by c=2, while NTS-NOTEARS gains. Robustness at the fixed decision budget is therefore method-specific rather than shared across the evaluated methods.

Lag localization. Conditional lag accuracy is comparatively stable across confounding strengths for several lag-capable methods, although some outputs deteriorate and NTS-NOTEARS improves (Appendix Figure 7). For these methods, confounding leaves lag placement among the true pairs entering their lag-evaluation sets largely intact. Complete four-strength trajectories are reported in Appendix Table 11.

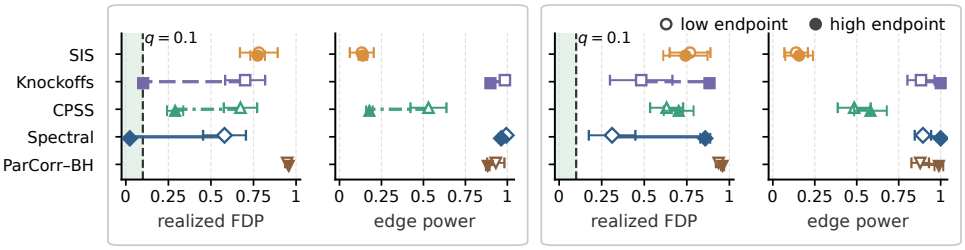
Error control. Reported selected sets exhibit a different response (Figure 3b): as confounding strengthens, mean realized FDP exceeds q=0.1 for several selectors while substantial edge power remains. Thus, degradation in error control need not arise from simply recovering fewer true edges.

Cross-layer contrasts. The within-layer responses do not move together (Figure 4). First, operating-point selection and lag localization are separable. SIS, PWGC, ParCorr–BH, Neural Granger, and CUTS have only modest precision at the shared $K{=}N$ budget, yet achieve conditional lag accuracy above 0.8 among the true pairs entering their lag-evaluation sets (Figure 4a). The gray bars show that these conditional denominators vary widely, so high lag accuracy over a small evaluated set should not be read as broad pair retrieval. Second, high pair AUPR does not imply low selected-set error (Figure 4b). Spectral deconfounding ranks pairs near-perfectly (AUPR 0.996) while its mean realized FDP is 0.200. Pair ranking therefore measures ordering quality, not error control. At this setting, comparable FDP inflation also appears under independent per-variable latent forcing for Spectral deconfounding and Knockoffs, whereas the SIS loss and NTS-NOTEARS gain in pair AUPR do not (Appendix Table 13).

Together, the within-layer responses and cross-layer contrasts show that ***performance under latent confounding has no single failure mode***. Responses vary across methods and evaluation layers, and strong performance in one layer need not carry over to another.

4.2 Effects of Observational Resources

To address **RQ2**, we examine how increasing the number of observed variables N or the record length T affects realized FDP and edge power under confounding. We first study how performance changes when varying N and T separately, and then use the (N, H, T) grid to identify the combinations at which error control is recovered.



(a) N: 20 → 120 (b) T: 300 → 4000

Figure 5: Resource-endpoint contrasts under strong confounding. (a) Mean realized FDP and edge power at N=20 and 120 (T=600). (b) The same metrics at T=300 and 4000 (N=20). Hollow and filled markers denote the lower and upper endpoints, and error bars show one SD.

Separate effects of cross-section and record length. The two observational resources produce qualitatively different, selector-dependent responses (Figure 5). Spectral deconfounding gives the clearest contrast: along N at T=600 its mean realized FDP falls by more than an order of magnitude while power stays near one (0.580 to 0.024), whereas along T at N=20 it moves to higher FDP. Knockoffs also lowers FDP sharply along N, whereas CPSS does so at a substantial cost in power and SIS and ParCorr–BH show little or no improvement. Systems at different N are generated independently rather than by adding nodes to one system, so the N path compares generated problem instances.

Record length at a fixed cross-section. At fixed N=20, the prespecified record-length family shows no hidden recovery between the endpoints. Across five selectors, five record lengths, and $c \in \{1, 2\}$, none of the 50 cells is recovered, 46 are off-target, and four SIS cells at c=1 are undecided (Appendix Figure 8). At this cross-section, longer records restore none of the five selectors over the evaluated range. At N=20, the rise in FDP with record length persists under full, half, and quarter loading for both spectral deconfounding and Knockoffs, although it weakens as loading becomes sparser (Appendix Table 14).

Cross-section and record length together. The crossed grid reveals a stronger result: longer records do not merely fail to substitute for a wider cross-section, they shift the spectral recovery threshold in N outward. On this grid, which classifies 150 cells per selector at c=2, the threshold for spectral deconfounding at a single latent driver widens roughly threefold across the evaluated record lengths and is right-censored at the longest (Figure 6 and Appendix Table 15). The shift is not a collapse of power: for both spectral deconfounding and Knockoffs, mean realized FDP rises with T at every N and driver count while the selected set grows and edge power ends higher than it began. At the longest record no selector recovers within the evaluated range of N.

Together, these results show that ***cross-section and record length have distinct, selector-dependent effects under confounding***. Here, wider cross-sections show lower realized FDP for several selectors, sometimes at a cost in power, whereas longer records show no comparable FDP reduction at fixed N=20.

4.3 Error-control recovery regimes and boundaries

To address **RQ3**, we ask what determines whether and where error control can recover.

Number of latent drivers. Under the benchmark parameterization, the recovery boundary for spectral deconfounding at T=600 moves outward as the number of latent drivers H increases, from $N^\star$=80 for $H \in \{1, 2\}$ to $N^\star$=120 for $H \geq 3$ (Figure 6). Because latent loadings are not energy-normalized across H, this comparison does not separate driver count from total latent influence. The other classified selectors show no comparable recovery pattern. This is consistent with the dense-confounding regime for which spectral deconfounding is designed. All reported $N^\star$ values are estimated at c=2, so their dependence on confounding strength is not characterized here.

The numerical boundary also depends on the prespecified multiplicity family. At H=1, the same run-level values yield $N^\star$=60 under the earlier 30-cell family but 80 under the crossed-grid 150-cell family, so the difference reflects correction width rather than the data (Appendix A.1). We therefore

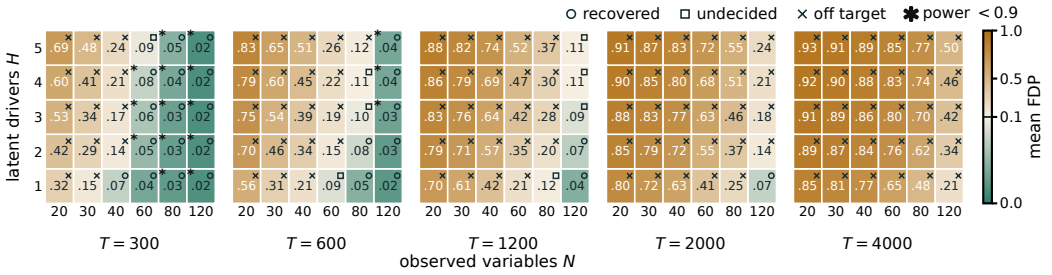


Figure 6: Mean realized FDP of Spectral deconfounding on the (N, H) grid under strong confounding, one panel per record length T. Cell fill and the printed value give mean realized FDP, and the markers give the frozen classifier verdict as drawn in the legend.

Table 3: Sensitivity analyses in the earlier (N, H) classification family, for spectral deconfounding at H=1, c=2, and T=600. Thresholds are thus on that family's scale, where the headline is $N^\star$=60.

Variation	Setting	Threshold $N^\star$	Power there
Earlier-family headline	$q = 0.1,\ L = 2$, full loading	$N^\star = 60$	1.00
Matched level q_M=q	0.05	$N^\star = 80$	0.99
	0.20	$N^\star = 60$	1.00
Loading breadth	0.25 of nodes	$N^\star = 120$	0.99
	0.50 of nodes	$N^\star = 120$	0.98
	1.00 of nodes	$N^\star = 60$	1.00
Maximum lag L	$L = 1$	$N^\star = 30$	0.99
	$L = 3$	$N^\star = 60$	0.98

report each $N^\star$ with its family, using the 150-cell family for crossed-grid results and the prespecified 30-cell family for Table 3.

Selector dependence. Recovery is strongly selector-specific on the crossed grid. Spectral deconfounding is the only selector with recovery thresholds, resolving 13 of its 25 evaluated slices (Appendix Table 15). The Knockoffs screen has recovered cells at the shortest record, but no scalar threshold because the larger N=120 cells are undecided. Its threshold is unresolved at T=600, and all cells are off-target thereafter. CPSS is off-target in all 150 evaluated cells (Appendix D.3). Thus, the recovery boundary is not shared across selectors.

Boundary sensitivity. The spectral-deconfounding boundary persists across the prespecified sensitivity analyses but shifts with the data-generating regime (Table 3). Restricting the latent loading from all observed variables to one half or one quarter moves $N^\star$ from 60 to 120, whereas reducing the maximum lag from L=2 to L=1 lowers it to 30. These thresholds are reached at near-unit power, and additional controls indicate that the shifts are not explained by edge prevalence alone (Appendix D.3).

Together, within the evaluated linear-Gaussian generator, these results show that ***recovery is not a fixed property of a method***. It is a regime property jointly determined by the selector, the driver-count and latent-energy regime, observational resources, and other generator conditions.

5 Conclusion

MirageBench provides a controlled benchmark for causal discovery under latent confounding, combining exact lagged-edge ground truth with output-aware Pole evaluation. Across the evaluated regimes, methods show no single failure mode under latent confounding: increasing the cross-section and the record length have distinct, selector-dependent effects on realized FDP and edge power, and error-control recovery is selector- and regime-dependent. These results shift attention from whether a method works to which structural claims remain reliable, and under what conditions. The benchmark is currently limited to controlled synthetic systems and finite grids. Formal recovery verdicts apply only to the prespecified lagged-selection families, other methods are evaluated descriptively, and FCI-family methods are excluded because the current ground truth does not define their latent-projected target. Future work should extend the recovery analysis beyond the sampled grid and the current dynamics, and test whether these regimes persist in observational systems with independently established causal structure.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on experimental design, interpretation of results, mathematical assistance and consistency checking. We did not use generative AI to generate experimental data or write mathematical proofs. Additionally, we used generative AI for language editing and assistance with code. We reviewed all AI-assisted work: numerical claims were verified against the underlying experimental outputs, and AI-assisted code and mathematical content were checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We release the MIRAGEBENCH generator, method adapters, POLE evaluation pipeline, configurations, and documentation needed to run the benchmark. The main generation and evaluation procedures are given in Section 3 and Algorithms 1–3, with complete methodological and implementation details in the appendices. Unless stated otherwise, experiments use 30 runs with metrics computed per run before aggregation. Appendix C records software versions, method settings, and per-result configurations. External implementation dependencies are documented in the released repository.

REFERENCES

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022. doi: 10.1613/JAIR.1.13428.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015. doi: 10.1214/15-AOS1337.

Lionel Barnett and Anil K. Seth. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, 2014. doi: 10.1016/j.jneumeth.2013.10.018.

Sumanta Basu, Xianqi Li, and George Michailidis. Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5):1207–1222, 2019. doi: 10.1109/TSP.2018.2887401.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Ben S. Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, 120 (1):387–422, 2005. doi: 10.1162/0033553053327452.

Kurt Butler, Daniel Waxman, and Petar M. Djurić. Tangent space causal inference: Leveraging vector fields for causal discovery in dynamical systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-3816.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

Luca Castri, Sariah Mghames, Marc Hanheide, and Nicola Bellotto. Enhancing causal discovery from robot sensor data in dynamic scenarios. In *Conference on Causal Learning and Reasoning (CLeaR)*, volume 213 of *Proceedings of Machine Learning Research*, pp. 243–258, 2023.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020.

Aditya Chaudhry, Pan Xu, and Quanquan Gu. Uncertainty assessment and false discovery rate control in high-dimensional Granger causal inference. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 684–693, 2017.

Zhengming Chen, Guangdong Sun, Fuhong Wang, Yewei Xia, Ruichu Cai, and Zhifeng Hao. Causal discovery from time-series discrete data in the presence of latent confounders. *Neural Networks*, 205:109426, 2027. doi: 10.1016/j.neunet.2026.109426.

Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS: Neural causal discovery from irregular time-series data. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. CUTS+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11525–11533, 2024a. doi: 10.1609/aaai.v38i10.29034.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *International Conference on Learning Representations (ICLR)*, 2024b.

Chien-Ming Chi, Yingying Fan, Ching-Kang Ing, and Jinchi Lv. High-dimensional knockoffs inference for time series data. *Journal of the American Statistical Association*, 120(551):1763–1774, 2025. doi: 10.1080/01621459.2024.2431344.

Ethan R. Deyle, M. Cyrus Maher, Ryan D. Hernandez, Sanjay Basu, and George Sugihara. Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences*, 113(46): 13081–13086, 2016. doi: 10.1073/pnas.1607747113.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911, 2008. doi: 10.1111/j.1467-9868.2008.00674.x.

Muhammad Hasan Ferdous, Emam Hossain, and Md Osman Gani. Timegraph: Synthetic benchmark datasets for robust time-series causal discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Datasets and Benchmarks Track*, pp. 5425–5435, 2025. doi: 10.1145/3711896.3737439.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12615–12625, 2020.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 1969. doi: 10.2307/1912791.

Wiebke Günther, Urmi Ninad, and Jakob Runge. Causal discovery for time series from multiple datasets with latent contexts. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research*, pp. 766–776, 2023.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, 50(3):1320–1347, 2022. doi: 10.1214/21-AOS2152.

Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. Causaldynamics: A large-scale benchmark for structural discovery of dynamical causal models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (56):1709–1731, 2010.

Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data, 2021.

Junning Li and Z. Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(17):475–514, 2009. URL https://www.jmlr.org/papers/v10/li09a.html.

Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of the 2018 ACM SIGKDD Workshop on Causal Discovery*, volume 92 of *Proceedings of Machine Learning Research*, pp. 23–47, 2018.

Ričards Marcinkevičs and Julia E. Vogt. Interpretable models for Granger causality using self-explaining neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Søren Wengel Mogensen, Karin Rathsman, and Per Nilsson. Causal discovery in a complex industrial system: A time series benchmark. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1218–1236, 2024.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. doi: 10.3390/make1010019.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1595–1605, 2020.

Panayiotis Panayiotou, Audrey Poinsot, Alessandro Leite, Nicolas Chesneau, Marc Schoenauer, and Özgür Şimşek. Causalprofiler: Generating synthetic benchmarks for rigorous and transparent evaluation of causal machine learning. *arXiv preprint arXiv:2511.22842*, 2025.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30 (4):962–1030, 2002. doi: 10.1214/aos/1031689015.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397, 2020.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1):2553, 2019a. doi: 10.1038/s41467-019-10105-3.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11):eaau4996, 2019b. doi: 10.1126/sciadv.aau4996.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, volume 123 of *Proceedings of Machine Learning Research*, pp. 110–120, 2020.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B*, 75(1):55–80, 2013. doi: 10.1111/j.1467-9868.2011.01034.x.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011. doi: 10.1016/j.neuroimage.2010.08.063.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000. ISBN 978-0-262-19440-2.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers – scaling up benchmarking of causal discovery for real-world time-series. In *International Conference on Learning Representations (ICLR)*, 2025.

Gideon Stein, Niklas Penzel, Tristan Piater, and Joachim Denzler. Tcd-arena: Assessing robustness of time series causal discovery methods against assumption violations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Eric V. Strobl, Peter L. Spirtes, and Shyam Visweswaran. Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology*, 10(5):1–37, 2019. doi: 10.1145/3351342.

Jie Sun, Dane Taylor, and Erik M. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015. doi: 10.1137/140956166.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge, 2021.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B. Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2022. doi: 10.1109/TPAMI.2021.3065601.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022. doi: 10.1017/eds.2022.11.

Raanan Yehezkel Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. From temporal to contemporaneous iterative causal discovery in the presence of latent confounders. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39939–39950, 2023.

Huiyang Yi, Xiaojian Shen, Yonggang Wu, Duxin Chen, He Wang, and Wenwu Yu. Causalcompass: Evaluating the robustness of time-series causal discovery in misspecified scenarios, 2026. arXiv:2602.07915.

Wanqi Zhou, Shuanghao Bai, Shujian Yu, Qibin Zhao, and Badong Chen. Jacobian regularizer-based neural Granger causality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61763–61782, 2024.

A PRELIMINARIES

Table 4 collects the symbols used throughout the appendix.

Table 4: Notation used in the appendix. The benchmark reference level q is distinct from any method-native nominal target q_M.

Symbol	Type	Meaning
N	scalar	number of observed variables
T	scalar	record length
L	scalar	maximum lag
$\mathbf{X} = \{x_{i,t}\}$	array	observed multivariate record
$\mathbf{\Theta} = \{\theta_{i,j,\tau}\}$	tensor	direct-edge coefficient tensor
$\mathbf{U} = \{u_{h,t}\}, \mathbf{B} = \{b_{i,h}\}$	array / matrix	latent drivers and their loadings
c, H	scalars	confounding strength and number of latent drivers
$\mathcal{C}$, $E^\star$	sets	candidate lagged-edge universe and direct-edge truth
$\mathcal{U}_{\text{pair}}$, $E^\star_{\text{pair}}$	sets	ordered cross-node pair universe and any-lag pair truth
$\widehat{E}_M$	set	lag-resolved set returned by configured method M
q	scalar	benchmark reference error level
$\overline{\text{FDP}}_M(z; T, c)$	scalar	mean realized FDP over runs at $z = (N, H)$
$\widehat{\mathcal{R}}_M(T, c; q)$	set	evaluated cells classified recovered
$N^\star_M(H; T, c, q)$	scalar	first sustained recovered N on the evaluated grid
$\mathcal{L}^\star_{ji}$	set	true lags of the ordered pair (j, i)
$\lambda_{M,s}$	parameters	configured parameters of method M at setting s, such as a native level or selection budget

Definition 3 (Confounding-induced false edge)**.** *For $c > 0$, a false reported edge $(j, \tau) \to i \in \widehat{E}_M \setminus E^\star$ is **confounding-induced** when its source and target share an injected latent driver.*

A reported false edge may be confounding-induced, when its endpoints share an injected latent driver, or may arise from other sources such as observation noise and finite-sample estimation error. Realized FDP counts both kinds, because the confounding-induced label is generator-level metadata rather than a second scoring rule. The label does not prove that the shared driver was the unique cause of an individual error, and it cannot be assigned from a real observed record alone, so every reported recovery result uses total FDP. Here, a false edge is a reported directed edge that is absent from the direct-edge ground truth. This differs from a confounded causal effect, where the estimate of a true causal relation is biased by confounding, and from a bidirected edge in acyclic directed mixed graphs (ADMGs) and partial ancestral graphs (PAGs) (Richardson & Spirtes, 2002), which marks latent common causation without asserting a direct edge.

A.1 RECOVERY CLASSIFICATION

For each selector-setting pair, the FDR estimand is $\mu = \mathbb{E}[\text{FDP}(\widehat{E}_M)]$, estimated by

$$\overline{\text{FDP}}_M(z; T, c) = \frac{1}{R} \sum_{r=1}^{R} \text{FDP}\left(\widehat{E}^{(r)}_M(z)\right). \tag{7}$$

The selector families, grids, and multiplicity rules were fixed in dated internal analysis plans before their corresponding results were inspected, not registered externally. The classifier tests $H_0 : \mu = q$ two-sided with a one-sample t-test on the $R = 30$ run-level values. After results inspection, the original $p = 0$ convention for zero-variance cells was corrected: their t statistic is undefined, so the amended rule assigns a conservative non-rejection placeholder $p = 1$ and leaves them undecided. A Holm correction at $\alpha = 0.05$ is applied separately within each selector's own prespecified set of evaluated cells, never jointly across selectors. Bootstrap intervals shown with some means are descriptive and do not determine the labels. The earlier (N, H) analysis fixed three selectors before

its results were seen, each corrected over its own 30-cell family at c=2; the separately prespecified record-length analysis covers five selectors, each corrected jointly over its 10 evaluated (T, c) settings. In the crossed grid at c=2, each of the same three selectors is corrected over all 150 evaluated (N, H, T) cells, so every comparison across record lengths is made inside one correction family.

Definition 4 (Evaluated settings and empirical recovery boundary)**.** *Let $\mathcal{N}_{\text{eval}}$ and $\mathcal{H}_{\text{eval}}$ be the finite evaluated values and $\mathcal{Z}_{\text{eval}} \subseteq \mathcal{N}_{\text{eval}} \times \mathcal{H}_{\text{eval}}$ the evaluated cells. Under the classifier above, define*

$$\widehat{\mathcal{R}}_M(T, c; q) = \{z \in \mathcal{Z}_{\text{eval}} : z \text{ is classified recovered}\}. \tag{8}$$

*Two cells are adjacent when they agree in one coordinate and are consecutive along the other evaluated coordinate. The **empirical recovery boundary** is the set of adjacent pairs with exactly one member in $\widehat{\mathcal{R}}_M(T, c; q)$.*

Definition 5 (Empirical recovery threshold)**.** *For fixed H, let $\mathcal{N}_{\text{eval}}(H) = \{N : (N, H) \in \mathcal{Z}_{\text{eval}}\}$. If a recovered cell is followed by a non-recovered cell at a larger evaluated N, we withhold the scalar threshold and report cell-wise classifications, even if a later recovered tail exists. Otherwise, when at least one cell is recovered, the grid threshold is*

$$N^{\star}_M(H; T, c, q) = \min\{N \in \mathcal{N}_{\text{eval}}(H) : (N', H) \text{ is classified recovered for every evaluated } N' \geq N\}. \tag{9}$$

If no cell is recovered, the threshold is recorded as $> N_{\max}(H)$, where $N_{\max}(H) = \max \mathcal{N}_{\text{eval}}(H)$, when every evaluated N at that H is off-target, and as unresolved when at least one cell is undecided. Neither status claims that no finite N can recover.

B Method Details of MirageBench

B.1 Controllable generator

The generator implements Eqs. 4 and 5. Each candidate direct edge is sampled independently, with configured bonuses for self-lags and designated spatial pairs; nonzero coefficients receive a random sign and a magnitude in the configured range. The lag matrices are assembled into a VAR companion matrix, and the entire coefficient tensor is rescaled until its spectral radius is below the configured cap. This scalar rescaling preserves the support $E^{\star}$. Each latent driver loads onto a configured group of $\min(g, N)$ observed variables, drawn without replacement, through signed weights and node-specific lags, with g=200 in the headline grids, giving full loading there. The configured edge-sampling probability, direct-effect magnitude range, coefficient clip, and spatial radius and bonus are listed in Table 9 rather than repeated in the algorithm. In the structural equation, the scalar c scales only the latent confounding term. Table 9 lists the shared generator defaults.

B.2 Evaluation pipeline

Algorithms 2 and 3 specify the common loop and the output-aware per-run evaluators. Each method receives the same observed record within a cell, while $E^{\star}$ and latent metadata remain evaluator-only. The four layer reports remain separate, and no composite score is formed.

B.3 Benchmark interface and evaluation tracks

This subsection makes explicit the interface rules underlying Algorithm 3 and Section 3.2.

Method-facing instance and output interface. A method receives the observed record $\mathbf{X}$, candidate universe $\mathcal{C}$, maximum lag, and permitted structural metadata. The direct-edge truth $E^{\star}$ and latent drivers are evaluator-only objects. A returned output may carry one or more of the following components: a selected *lagged set*, lag-resolved evidence as scores or p-values, a pair-level output as a score, matrix, or selected set, or a structured graph such as a PAG. Structured graphs are declared by the schema but are not scored because the corresponding latent-projected truth has not been derived for this generator. Layer membership follows the capabilities carried by the returned object, not a mutually exclusive method category.

Algorithm 1 MIRAGEBENCH generator.

Input: registered generator setting s and run index r, with the recorded random state for (s, r). Setting s specifies $N, T, T_{\text{burn}}, L, c, H$, group-size cap g, ρ, $[b_{\text{lo}}, b_{\text{hi}}]$, σ_ε, alignment, and spectral cap $r^\star$, together with the registered edge-sampling, coefficient-range, clipping, and spatial parameters of Table 9

// Step 1: spatial layout.

Place N nodes on the synthetic unit square and form the spatial adjacency $\mathcal{A}$

// Step 2: direct lagged graph.

Sample the sparse direct-edge tensor $\mathbf{\Theta}$ over (i, j, τ), $\tau \le L$, boosting pairs chosen by *alignment* relative to $\mathcal{A}$, then rescale $\mathbf{\Theta}$ to a VAR companion radius $< r^\star$

Set the ground-truth edges $E^\star = \{(j, \tau) \to i : \theta_{i,j,\tau} \neq 0\}$

// Step 3: latent drivers.

Initialize $\mathbf{U} \leftarrow \mathbf{0}_{H \times (T + T_{\text{burn}} + L)}$ and empty group, loading, and node-lag metadata

if $c > 0$ **and** $H > 0$ **then**

  for $h = 1, \ldots, H$ **do**

    Draw the AR(1) driver $u_{h,t} = \rho\, u_{h,t-1} + \eta_{h,t}$, $\eta_{h,t} \sim \mathcal{N}(0, 1)$, and a node group G_h of $\min(g, N)$ observed variables drawn without replacement

    for $i \in G_h$ **do**

      Draw $b_{i,h} \sim \pm\,\text{Unif}[b_{\text{lo}}, b_{\text{hi}}]$ and $\ell_{i,h} \in [L]$

    end for

    Set $b_{i,h} = 0$ for $i \notin G_h$

  end for

else

  Make no latent-driver, group, loading, or node-lag random draws

end if

// Step 4: time series rollout.

Initialize the observed array to zero and draw independent Gaussian values for $x_{i,t}$, $t \in [L]$

for $t = L + 1, \ldots, L + T_{\text{burn}} + T$ **do**

  for $i = 1, \ldots, N$ **do**

    $x_{i,t} = \sum_{\tau=1}^{L} \sum_{j=1}^{N} \theta_{i,j,\tau}\, x_{j,t-\tau} + c \sum_{h=1}^{H} b_{i,h}\, u_{h,t-\ell_{i,h}} + \varepsilon_{i,t}$, $\varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$

  end for

end for

// Step 5: output separation.

Extract the final T steps as $\mathbf{X}$ after the L initialization and T_{burn} burn-in steps

Package $\mathbf{X}$ as method-facing data without $E^\star$, $\mathbf{\Theta}$, or latent metadata

return the method-facing data and evaluator metadata to the benchmark runner

Random-ranking reference for pair AUPR. Pair AUPR is tie-averaged average precision, obtained by averaging exactly over all orders within each score-tie group. For scored universe $\mathcal{U}$ and positive set P, the exact expected average precision under random ranking is

$$\text{AP}_{\text{rand}}(P, \mathcal{U}) = \frac{1}{g}\left[H_g + \frac{p-1}{g-1}(g - H_g)\right], \qquad g = |\mathcal{U}|, \quad p = |P \cap \mathcal{U}|, \tag{10}$$

where H_g is the gth harmonic number. This finite-universe expectation, rather than prevalence p/g, is the random-ranking reference for average precision. It is an expected value under random ranking, not a lower bound: a single random ranking, or a poor method, can fall below it.

Self-pairs. The common pair-level target is restricted to cross-node relations, while self-lags remain in the lag-resolved candidate universe. Because autoregressive self-links are typically much stronger than cross-node relations, including them would change the composition of the pair-level task, so Table 5 reports the corresponding sensitivity rather than treating the restriction as innocuous. Including self-pairs changes pair AUPR only modestly for most methods, which indicates that the cross-node comparisons are not an artifact of the restriction.

Method contracts. Table 6 records each evaluated method's group, field-importance tier, adapter output, admitted POLE layers, and membership in the earlier frozen three-selector (N, H) classification family. A dash marks an unsupported layer rather than a zero-valued result, so a method without a configured lagged selected set receives no E-layer FDP, power, or recovery verdict. E-layer membership is recorded from the set a method's adapter actually returns, and that set has three provenances. It is native for the screens, the knockoff constructions, and the constraint-based methods; it is formed by the benchmark's fixed multiple-testing rule for the Granger variants, the PCMCI variants, and the ParCorr–BH comparator; and for GVAR, TCDF, NTS-NOTEARS, and

Algorithm 2 MIRAGEBENCH evaluation loop.

Input: prespecified evaluation settings $\mathcal{S}$ and their run indices, configured methods and adapters $\mathcal{M}$, frozen projection and evaluation rules Π, and the documented classification protocol $\mathcal{P}$ where applicable

Initialise: per-run reports $\mathcal{D} \leftarrow \varnothing$

// Stage 1: per-run evaluation.

for $s \in \mathcal{S}$ and each required run index r for s **do**

  Retrieve or prepare once the fixed record $\mathbf{X}_{s,r}$ and evaluator-side truth $E^{\star}_{s,r}$ under the fixed lineage

  Derive $\mathcal{C}_s$, $\mathcal{U}_s$, and $E^{\star}_{\mathrm{pair},s,r}$ from s, $E^{\star}_{s,r}$, and the frozen rules Π

  for each configured method $M \in \mathcal{M}$ **do**

    Make the same $\mathbf{X}_{s,r}$ available to M

    $I_{M,s,r} \leftarrow$ construct the admissible method-facing instance from $\mathbf{X}_{s,r}$ and allowed metadata ▷ truth withheld

    $O_{M,s,r} \leftarrow$ call the configured adapter, its method-specific preprocessing, and M on $I_{M,s,r}$

    Preserve every returned score, selected-set, and lag component that is present without inventing absent components

    $D_{M,s,r} \leftarrow \mathrm{EvaluateLayers}(O_{M,s,r}, E^{\star}_{s,r}, E^{\star}_{\mathrm{pair},s,r}, \mathcal{C}_s, \mathcal{U}_s, \Pi, \lambda_{M,s})$ using Algorithm 3

    Store $(s, r, M, O_{M,s,r}, D_{M,s,r})$ in $\mathcal{D}$

  end for

end for

// Stage 2: aggregate run-level metrics.

Aggregate each per-run metric in $\mathcal{D}$ over the required runs, taking means and standard deviations of the per-run values rather than recomputing nonlinear metrics such as F1 from aggregated inputs

// Stage 3: classify eligible families.

if $\mathcal{P}$ registers the experiment, settings, lineage, run counts, and multiplicity family **then**

  Assemble the complete prespecified family. Retain every absent or incomplete cell in the family at $p = 1$, so the family size and every other cell's correction are unchanged

  Apply the frozen batch classification in $\mathcal{P}$ to the family ▷ incomplete cells cannot recover

end if

return descriptive summaries and any classifications admitted by $\mathcal{P}$

Table 5: Self-pair sensitivity of pair AUPR at c=2, N=40, and H=1 over 30 runs. "Including self" evaluates the joint pair universe with self-pairs included, "Change" is taken against the cross-node value in Table 11, and "Self pairs only" evaluates self-pairs alone.

Method	Including self	Change	Self pairs only
SIS	0.232	+0.012	0.539
Knockoffs	0.980	+0.001	0.996
CPSS	0.536	+0.012	0.712
Spectral deconf.	0.996	+0.000	1.000
ParCorr–BH	0.266	+0.011	0.612
PCMCI	0.927	+0.006	0.990
F-PCMCI	0.891	+0.002	0.922
oCSE	0.909	−0.076	0.291
VAR	0.998	+0.000	1.000
MVGC	0.996	+0.000	1.000
PWGC	0.276	+0.012	0.608
VARLiNGAM	0.998	+0.000	1.000
Neural Granger	0.329	+0.013	0.623
GVAR	0.994	+0.001	1.000
TCDF	0.665	+0.009	0.893
CUTS+	0.758	+0.010	0.916
CUTS	0.412	+0.013	0.747
NTS-NOTEARS	0.889	+0.009	0.953
DYNOTEARS	0.915	+0.007	0.987
TSCI	0.112	−0.005	0.291

DYNOTEARS it rests on a benchmark-side rule applied to a continuous output, a weight threshold for the two NOTEARS variants and a lag attached by the evaluator to a pair-level selection for the other two. The last group's error-control values therefore depend on a benchmark choice as well as on the method. Methods whose native output is a structured graph are absent from the table and receive no layer scores at all. LPCMCI is the evaluated instance of this case: its partial ancestral graph asserts ancestral direction together with adjacency, the benchmark truth is one-hop, and no latent-projected or ancestral ground truth is defined here to score it against (Appendix D.4).

Algorithm 3 Per-run POLE evaluation.

Input: returned object O, truth $E^\star$ and $E^\star_{\text{pair}}$, universes $\mathcal{C}$ and $\mathcal{U}$, frozen score, set, operating-point, and lag rules Π, and configured parameters including $K{=}N$

Initialise: report $\leftarrow \varnothing$ and $\mathbf{s}_{\text{pair}} \leftarrow \text{NA}$ ▷ NA: unsupported, not zero

// P. Pair ranking.

if O carries a legal per-pair or per-candidate score, not only a set-membership indicator **then**

  $\mathbf{s}_{\text{pair}} \leftarrow \text{ProjectScores}(O, \mathcal{U}; \Pi)$ using $\max_\tau s$, $-\log\max(\min_\tau p, 10^{-300})$, clipping p below at 10^{-300}, or the declared matrix identity $S[j,i] = s(j \to i)$. Fill unanswered pairs with O's worst score

  Add tie-averaged average precision, reported as pair AUPR, against $E^\star_{\text{pair}}$, the expected random-ranking average precision of Eq. 10, and prevalence as separate quantities on $\mathcal{U}$

end if

// O. Operating points.

if $\mathbf{s}_{\text{pair}} \neq \text{NA}$ **then**

  Evaluate the top $K{=}N$ projected pairs, sharing a score tie at the cutoff by its expected fraction

  Add pair precision and recall for this run at the shared budget

end if

if O carries a reported selected set, including an empty set **then**

  $Q_{\text{reported}} \leftarrow \text{ProjectSet}(O, \mathcal{U}; \Pi)$ by selection at any lag with pair deduplication, or by the pair-set identity, then restrict to $\mathcal{U}$

  Add pair precision, recall, and F1 for this run at Q_{reported}

end if

// L. Lag localization.

if O carries lag information **then**

  if O carries a reported selected set, including an empty set **then**

    $Q_L \leftarrow \text{ProjectSet}(O, \mathcal{U}; \Pi)$

  else if O carries a legal score **then**

    $Q_L \leftarrow$ the top-K pair set from $\text{ProjectScores}(O, \mathcal{U}; \Pi)$ at $K{=}N$, breaking a tie at the cutoff by the universe's fixed enumeration order

  else

    $Q_L \leftarrow \text{NA}$

  end if

  if $Q_L \neq \text{NA}$ **then**

    For each $(j,i) \in Q_L$, let A_{ji} contain all greatest-strength or smallest-p lags when per-lag values exist, and otherwise the reported lag set or $\varnothing$

    $\mathcal{T} \leftarrow Q_L \cap E^\star_{\text{pair}}$

    if $|\mathcal{T}| > 0$ **then**

      Set $a_{ji} \leftarrow |A_{ji} \cap \mathcal{L}^\star_{ji}|/|A_{ji}|$ when $|A_{ji}| > 0$, and $a_{ji} \leftarrow 0$ otherwise, where $\mathcal{L}^\star_{ji} = \{\tau : (j,\tau) \to i \in E^\star\}$ is the set of true lags for that pair

      Add $|\mathcal{T}|^{-1} \sum_{(j,i)\in\mathcal{T}} a_{ji}$ and denominator $|\mathcal{T}|$ ▷ ties share credit

    else

      Add a missing lag-localization value and denominator 0

    end if

  else

    Add a missing lag-localization value

  end if

end if

// E. Error control.

if O carries a selected lagged set $\widehat{E} \subseteq \mathcal{C}$ satisfying the E-layer contract **then**

  Add FDP $= |\widehat{E} \setminus E^\star|/\max(1, |\widehat{E}|)$ ▷ empty set: FDP 0

  Add edge power $|\widehat{E} \cap E^\star|/|E^\star|$ when $|E^\star| > 0$, and NA otherwise

end if

return report

C EXPERIMENTAL SETTINGS

C.1 METHOD CONFIGURATIONS AND NATIVE TARGETS

The evaluated methods divide into those implemented in this benchmark from their published procedures and those run from their authors' released code.

- **Implemented here.** The marginal screen (Fan & Lv, 2008), the knockoff screens (Barber & Candès, 2015; Candès et al., 2018), CPSS (Shah & Samworth, 2013), spectral deconfounding (Ćevid et al., 2020; Guo et al., 2022), VAR and pairwise Granger causality (Granger, 1969),

Table 6: Method outputs and POLE-layer admission under the benchmark adapters. **Class.** indicates membership in the earlier frozen three-selector (N, H) recovery-classification family, and † denotes methods evaluated in the benchmark but not included in the frozen method set of Figure 1. Spectral, NGC, and NG-cLSTM abbreviate Spectral deconfounding, Neural Granger, and Neural Granger (cLSTM).

Method	Group	Tier	Adapter output	Layers	Class.
SIS	Statistical	A	lagged set, scored	POLE	–
Knockoffs		A	lagged set, scored	POLE	Yes
CPSS		A	lagged set, sel. prob.	POLE	Yes
Spectral deconf.		A	lagged set, scored	POLE	Yes
VAR		A	lagged coefficients	POL–	–
PWGC		A	pairwise F	POLE	–
MVGC		A	conditional F	POLE	–
VARLiNGAM		A	lagged coefficients	POL–	–
PCMCI	Constraint	A	lagged set, p	POLE	–
ParCorr–BH		A	lagged set, p	POLE	–
F-PCMCI		A	lagged set, p	POLE	–
oCSE		A	lagged set, entropy	POLE	–
Neural Granger	Neural	A	lagged scores, support	POL–	–
Neural Granger (cLSTM)†		B	pair matrix	PO––	–
GVAR		B	lagged scores	POLE	–
TCDF		A	pair, one delay	POLE	–
CUTS		B	lagged matrix	POL–	–
CUTS+		A	pair matrix	PO––	–
JRNGC†		A	shared-model Jacobian	POL–	–
NTS-NOTEARS		A	lagged weights	POLE	–
DYNOTEARS	Other	A	lagged weights	POLE	–
TSCI		A	pair score	PO––	–

multivariate Granger causality (Barnett & Seth, 2014), VARLiNGAM (Hyvärinen et al., 2010), and the ParCorr–BH comparator, under the settings of Table 7.

- **Run from the authors' released implementations, unmodified.** PCMCI (Runge et al., 2019b), F-PCMCI (Castri et al., 2023), and the LPCMCI diagnostic (Gerhardus & Runge, 2020) through Tigramite[1]; the TSKI sampler (Chi et al., 2025) through `knockpy`[2]; oCSE (Sun et al., 2015) through its maintainer's package; and Neural Granger in its cMLP and cLSTM variants (Tank et al., 2022), GVAR (Marcinkevičs & Vogt, 2021), TCDF (Nauta et al., 2019), CUTS (Cheng et al., 2023), CUTS+ (Cheng et al., 2024a), JRNGC (Zhou et al., 2024), NTS-NOTEARS (Sun et al., 2021), DYNOTEARS (Pamfil et al., 2020), and TSCI (Butler et al., 2024) from their own repositories.

Each released implementation runs in the environment its own pins require, and the released repository records the source repository, commit, and licence of every one.

The LPCMCI diagnostic records Tigramite 5.2.6.7, while the earlier runs of the other two predate a correction to the version probe and record only that the package was present. Every knockoff result on the (N, H) family and the crossed grid uses the configured second-order Gaussian model-X construction with an equicorrelated s and a single block, pinned by configuration rather than selected at runtime. The TSKI diagnostic alone uses `knockpy` 1.3.2's Gaussian semidefinite-program sampler, averaging its statistic over five block subsamples at block fraction 0.3. Runs use NumPy 1.26.4, SciPy 1.11.4, and scikit-learn 1.2.2; recorded CPython patch versions vary within 3.11 but use the same numerical stack. Where a method exposes a nominal level it receives a native target of $q_M = 0.1$ unless stated otherwise. Other methods use their configured threshold or selection budget.

Native targets and statistical guarantees. Table 8 records the native statistical target of representative evaluated and adjacent methods together with the condition under which that target holds. Pair AUPR, pair F1, lag accuracy, realized FDP, and a method's native guarantee are distinct claims, and a native guarantee such as FDR control for multiple testing (Benjamini & Hochberg, 1995), variable selection (Barber & Candès, 2015; Candès et al., 2018; Chi et al., 2025), or Granger and PC-style graph recovery (Chaudhry et al., 2017; Li & Wang, 2009; Strobl et al., 2019) attaches to its own

[1] https://github.com/jakobrunge/tigramite

[2] https://github.com/amspector100/knockpy

Table 7: Settings of the principal set-returning methods.

Method	Setting
Marginal screen (SIS)	Top 0.02 fraction of the lagged cross-correlation ranking
Knockoff, headline	$q_M = 0.1$, lasso statistic, knockoff+ aggregation, one full block, Gaussian construction
Knockoff, TSKI variant	Time-series construction with 5 blocks, sampling fraction 0.3, and `knockpy`'s semidefinite-program Gaussian sampler
CPSS	50 subsamples, frequency threshold $\pi_{\mathrm{thr}} = 0.6$, selection budget $q_{\mathrm{select}} = 3$
Spectral deconfounder	$q_M = 0.1$, singular-value trim, Benjamini–Hochberg over per-edge p-values
ParCorr–BH	Partial-correlation tests conditioning on the target's own lags at $\alpha = 0.1$, followed by Benjamini–Hochberg

assumptions and error object, so it does not transfer automatically to direct lagged-edge FDP under latent confounding.

Table 8: Native targets and assumptions of representative evaluated and adjacent methods. The last column records how the native guarantee relates to the benchmark's direct-edge FDP target. BH: Benjamini–Hochberg.

Method	Group (family)	Output	Native target	Required condition	Relation to benchmark FDP target
Marginal / SIS (Fan & Lv, 2008)	Statistical (marginal screening)	ranked edges	Screening; sure-screening, not FDR	Marginal-correlation faithfulness; sparsity	No FDP guarantee
Fixed-X / Model-X / TSKI knockoffs (Barber & Candès, 2015; Candès et al., 2018; Chi et al., 2025)	Statistical (Model-X)	directed lagged edges	FDR, finite-sample or asymptotic by regime	Fixed design; known or parametric feature law; TSKI assumptions	FDR guarantee under different assumptions
CPSS (Shah & Samworth, 2013)	Statistical (stability selection)	directed lagged edges	Bounds selections of low-selection-probability covariates, not FDR	CPSS conditions (exchangeability-free base result)	Bounds low-probability selections, not FDP
Spectral deconfounder (Guo et al., 2022; Ćevid et al., 2020)	Statistical (spectral regression)	directed lagged edges	Sparse estimation or coefficient inference, not edge-set FDR	Dense hidden confounding	No edge-set FDR guarantee
ParCorr–BH comparator	Constraint group (rule-based comparator implemented here)	directed lagged edges	Discoveries from BH-adjusted partial-correlation tests, not an edge-set FDR guarantee	Causal sufficiency; valid partial-correlation p-values	No edge-set FDR guarantee
CUTS+ (Cheng et al., 2024a)	Neural (message-passing Granger)	pair-level summary graph	Continuous score per node pair, no lag index	No nominal FDR target; irregular sampling is its regime	Different scored object

C.2 Generator defaults and experiment matrix

Table 9 defines the generator defaults once, and Table 10 records what each experiment family varies and fixes relative to them.

Table 9: Shared generator defaults, used by every experiment unless Table 10 states otherwise.

Parameter	Default	Parameter	Default
Maximum lag L	2	Latent AR coefficient ρ	0.5
Burn-in	300	Latent-loading magnitude	$[0.3, 0.6]$
Direct-edge sampling probability	0.01	Spatial radius	0.30
Direct-effect magnitude	$[0.25, 0.45]$	Alignment	misaligned
Coefficient clip	0.5	Observation-noise standard deviation	1.0
Spectral-radius cap	0.90	Spatial edge bonus	0.10

C.3 Run construction and execution

Pairing across the varied axes. Within a run, the layout, $\mathbf{\Theta}$, and $E^\star$ are drawn before the latent drivers and loadings and held fixed across confounding strengths, and the positive strengths also share the latent-driver draws and the innovation noise, so they differ only through c, whereas the $c = 0$ reference shares the layout and $\mathbf{\Theta}$ but uses an independent noise realization. Different values of N are independently generated systems that use the same set of run indices, not nested additions to one graph. Across H, the latent-driver draws are nested at a common seed, so previously instantiated

Table 10: Experiment matrix relative to Table 9. Latent drivers are unobserved, with full loading unless a row varies it, and the benchmark reference level is q=0.1. Main-text and appendix figures report subsets of these experiment families.

Experiment	Varied axis	Fixed setting	Runs
Confounding response	$c \in \{0, 0.5, 1, 2\}$	$N = 40$, $H = 1$, $T = 600$	30
Earlier (N, H) family	$N \in \{20, 30, 40, 60, 80, 120\}$, $H \in \{1, \ldots, 5\}$	$c = 2$, $T = 600$	30
Record-length trajectory	$T \in \{300, 600, 1200, 2000, 4000\}$, $c \in \{1, 2\}$	$N = 20$, $H = 1$	30
Crossed-grid family	$N \in \{20, \ldots, 120\} \times H \in \{1, \ldots, 5\} \times T \in \{300, \ldots, 4000\}$	$c = 2$	30
Prespecified sensitivity	matched level q_M=$q \in \{0.05, 0.2\}$, loading breadth $\in \{0.25, 0.5\}$, or $L \in \{1, 3\}$, each over $N \in \{20, \ldots, 120\}$	$c = 2$, $H = 1$, $T = 600$; the level variation also spans $H \in \{1, \ldots, 5\}$	30
Truth prevalence	$N \in \{10, 20, 30, 40, 60, 80, 120\}$, generator only	defaults, no selector run	10
TSKI knockoff diagnostic	$c \in \{0, 2\}$	$N = 20$, $H = 1$, $T = 600$	5

drivers keep their series, groups, loadings, and node lags while each added driver appends a newly drawn loading, and the layout and $\mathbf{\Theta}$ stay fixed while the observed record is regenerated. Different values of T are separately rolled-out records with a fixed layout and $\mathbf{\Theta}$, not prefixes of one trajectory.

Execution of the crossed grid. The crossed grid was executed on two machines, a local workstation and a shared multi-core compute server, under one frozen configuration and one recorded per-run digest, and all 13,500 method runs completed with no failed run. **Protocol amendment.** On 14 September 2026, before any evaluation metric was inspected, the per-run wall-clock cap was raised from 7,200 to 28,800 seconds, together with the number of concurrent jobs, after runtime diagnostics at the largest cross-section showed that the initial cap could not be met at the longer records. The amendment and the concurrency setting were recorded before the affected jobs were rerun. Two runs that exceeded the original cap were quarantined with their diagnosis, a harness-level output-buffer deadlock rather than a method failure, and were rerun under the corrected harness with the same configuration and run indices. All 13,500 planned method runs subsequently completed and passed lineage and configuration validation before any metric was aggregated, each carrying digests of its record, its ground truth, its configuration, its sources, and its environment.

D More Experimental Results

The subsections follow the three research questions of Section 4, and Appendix D.4 adds one post-hoc diagnostic outside the prespecified families.

D.1 RQ1: Responses to latent confounding

Table 11 gives the complete pair-ranking and lag-localization trajectories behind Figure 1a and Section 4.1. Figure 7 gives the matched clean-to-strong changes in fixed-budget pair recall and conditional lag accuracy. The two changes are compared each with itself because their feasible maxima differ: at N=40 the top-K budget permits only 40 selected pairs, while the 30 instances contain 79–115 true pairs, giving an average Recall@K ceiling of 0.425. Table 12 gives the operating-point and error-control values behind Figure 1b.

Lag credit is conditional on the evaluated pair set Q_L, the reported selected pairs when a set is available and the shared top-K pair set otherwise, so its denominator differs by method: SIS 17.3, Knockoffs 93.5, CPSS 44.5, Spectral deconf. 94.6, ParCorr–BH 93.1 true directed pairs in the evaluated set per run on average. For reference, the finite-universe random-ranking average precision for P pair AUPR is 0.0651, and the E-layer target is $1 - q = 0.9$; these are metric-specific references rather than radar-wide baselines.

Shared versus private latent forcing. Table 13 reports a prespecified control at the RQ1 setting that keeps the law of the latent forcing but removes its sharing across variables.

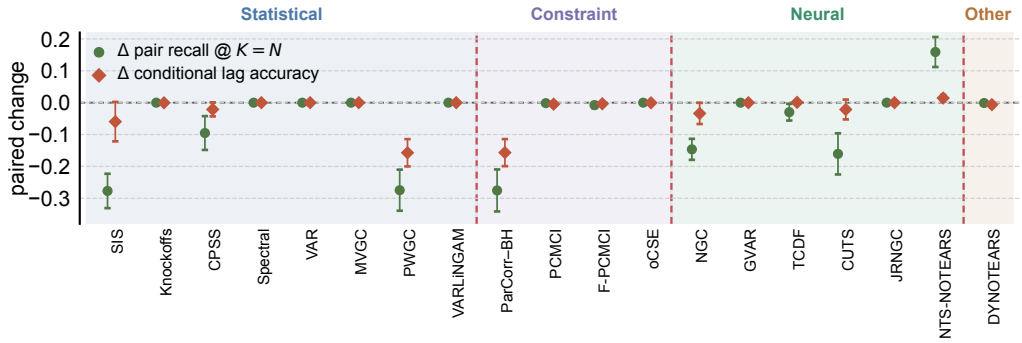


Figure 7: Change from $c{=}0$ to $c{=}2$ in pair recall at $K{=}N$ and in conditional lag accuracy at $N{=}40$, with one-SD bars over 30 runs. NGC: Neural Granger.

Table 11: Pair AUPR and conditional lag accuracy across confounding strengths at $N{=}40$, $H{=}1$, and $T{=}600$, as mean $\pm$ SD over 30 runs. Shared-budget Recall@N is the fraction of pair truth in the top N, a dash marks a method without a lag axis, and † denotes methods evaluated in the benchmark but not included in the frozen method set of Figure 1.

	Pair ranking (P)					Lag localization (L)			
Method	$c{=}0$	$c{=}0.5$	$c{=}1$	$c{=}2$	Recall@N $c{=}2$	$c{=}0$	$c{=}0.5$	$c{=}1$	$c{=}2$
SIS	0.977 ± 0.028	0.968 ± 0.033	0.756 ± 0.074	0.220 ± 0.065	0.146	1.000 ± 0.000	0.999 ± 0.005	0.995 ± 0.009	0.941 ± 0.062
Knockoffs	1.000 ± 0.001	1.000 ± 0.001	0.997 ± 0.004	0.979 ± 0.016	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.002	1.000 ± 0.002
CPSS	0.934 ± 0.030	0.927 ± 0.032	0.870 ± 0.042	0.524 ± 0.083	0.330	1.000 ± 0.000	1.000 ± 0.000	0.997 ± 0.006	0.979 ± 0.022
Spectral deconf.	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.001	0.996 ± 0.004	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
ParCorr–BH	0.984 ± 0.019	0.977 ± 0.014	0.835 ± 0.048	0.254 ± 0.076	0.149	0.999 ± 0.004	0.999 ± 0.003	0.975 ± 0.018	0.843 ± 0.042
PCMCI	1.000 ± 0.000	1.000 ± 0.001	0.990 ± 0.007	0.921 ± 0.022	0.424	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.002	0.995 ± 0.008
F-PCMCI	1.000 ± 0.002	1.000 ± 0.000	0.980 ± 0.009	0.889 ± 0.027	0.417	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.003	0.997 ± 0.007
oCSE	1.000 ± 0.000	1.000 ± 0.001	0.998 ± 0.004	0.984 ± 0.012	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.002	1.000 ± 0.002
VAR	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.002	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
MVGC	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.001	0.996 ± 0.003	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
PWGC	0.983 ± 0.020	0.978 ± 0.014	0.852 ± 0.040	0.264 ± 0.070	0.150	0.999 ± 0.004	0.999 ± 0.003	0.975 ± 0.018	0.842 ± 0.043
VARLiNGAM	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.002	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Neural Granger	0.536 ± 0.044	0.525 ± 0.039	0.476 ± 0.038	0.316 ± 0.040	0.227	0.988 ± 0.015	0.984 ± 0.021	0.982 ± 0.023	0.954 ± 0.040
GVAR	1.000 ± 0.001	0.999 ± 0.001	0.999 ± 0.002	0.993 ± 0.005	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
TCDF	0.796 ± 0.070	0.795 ± 0.065	0.743 ± 0.068	0.656 ± 0.073	0.374	0.999 ± 0.004	1.000 ± 0.000	0.999 ± 0.003	1.000 ± 0.000
CUTS+	0.988 ± 0.011	0.981 ± 0.011	0.951 ± 0.018	0.748 ± 0.072	0.382	–	–	–	–
CUTS	0.891 ± 0.034	0.866 ± 0.035	0.716 ± 0.049	0.399 ± 0.069	0.259	0.999 ± 0.005	0.998 ± 0.007	0.995 ± 0.014	0.978 ± 0.030
NTS-NOTEARS	0.484 ± 0.084	0.501 ± 0.078	0.603 ± 0.083	0.880 ± 0.053	0.398	0.985 ± 0.010	0.980 ± 0.016	0.984 ± 0.011	0.999 ± 0.003
DYNOTEARS	1.000 ± 0.000	0.999 ± 0.003	0.987 ± 0.012	0.908 ± 0.037	0.424	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.003	0.994 ± 0.008
TSCI	0.223 ± 0.042	0.233 ± 0.044	0.211 ± 0.039	0.116 ± 0.023	0.076	–	–	–	–
JRNGC†	1.000 ± 0.000	1.000 ± 0.001	0.999 ± 0.002	0.993 ± 0.006	0.425	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
Neural Granger (cLSTM)†	0.133 ± 0.026	0.134 ± 0.024	0.126 ± 0.023	0.097 ± 0.016	0.059	–	–	–	–

D.2 RQ2: Resource-dependent error-control recovery

Figure 8 retains every evaluated record length at $c{=}1$ and $c{=}2$ behind the endpoint view of Figure 5b, and marks the frozen record-length classification on each FDP point.

Loading breadth and record length. Table 14 reports a prespecified control at $N{=}20$ that repeats the record-length comparison with the latent driver loading on all, half, or a quarter of the observed variables. Since each driven variable keeps the same loading law, breadth and total latent energy change together, and this control does not separate them.

Crossed-grid family. Table 15 gives $N^\star(H,T)$ or its status for the three classified selectors in the crossed-grid family at $c{=}2$, and Figure 9 is its cell-level record, with 30 runs in each of the 150 cells per selector. A cell marked > 120 is right-censored: every evaluated cross-section was tested and rejected above q, so a threshold, if one exists, lies past the grid. A cell marked unresolved makes no such statement, because the cell that would carry the threshold is one the test could not decide. At $T{=}300$, the Knockoffs $N{=}80$ cells are recovered, but the $N{=}120$ cells are undecided, so no scalar threshold is reported. The recovered $N{=}80$ cells have mean FDP 0.023 to 0.049 and mean power 0.34 to 0.73.

Table 12: Reported-set O-layer and E-layer values behind Figure 1b at c=2, as mean ± SD over 30 runs, with the O-layer columns evaluated at each method's reported selected set rather than the shared budget. The P and L spokes are the c=2 columns of Table 11.

Selector	O precision	O recall	O pair F1	E power	E 1−FDP
SIS	0.304 ± 0.092	0.182 ± 0.048	0.227 ± 0.062	0.171 ± 0.047	0.291 ± 0.082
Knockoffs	0.622 ± 0.124	0.984 ± 0.016	0.755 ± 0.102	0.981 ± 0.016	0.642 ± 0.121
CPSS	0.588 ± 0.091	0.471 ± 0.066	0.521 ± 0.071	0.450 ± 0.058	0.594 ± 0.088
Spectral deconf.	0.786 ± 0.100	0.996 ± 0.005	0.875 ± 0.068	0.996 ± 0.005	0.800 ± 0.095
ParCorr–BH	0.072 ± 0.008	0.981 ± 0.014	0.133 ± 0.013	0.928 ± 0.028	0.050 ± 0.006

Table 13: Shared versus private latent forcing at N=40, H=1, T=600, and c=2, with means and SD over 30 runs. Private forcing replaces the shared latent term with independent per-variable processes of the same law. The comparison is descriptive and carries no verdict, and its clean arm shares a noise realization with the forced arms, so it is not comparable cell by cell with the main-text c=0 reference.

	Pair AUPR			Realized FDP		
Method	Clean	Shared	Private	Clean	Shared	Private
Spectral deconfounding	1.000 ± 0.000	0.996 ± 0.003	1.000 ± 0.000	0.030 ± 0.017	0.204 ± 0.090	0.272 ± 0.027
Knockoffs	0.999 ± 0.003	0.981 ± 0.011	0.999 ± 0.002	0.095 ± 0.041	0.381 ± 0.134	0.333 ± 0.045
SIS	0.978 ± 0.023	0.238 ± 0.054	0.952 ± 0.055	0.015 ± 0.038	0.693 ± 0.068	0.139 ± 0.086
NTS-NOTEARS	0.470 ± 0.070	0.875 ± 0.051	0.501 ± 0.088	0.868 ± 0.020	0.486 ± 0.090	0.859 ± 0.024

Two families at the shared record length. The crossed-grid family re-measures T=600, where the earlier (N, H) family of Table 16 was classified, with an independent set of runs. The two are classified as disjoint families under one rule and are never pooled, with one 30-cell Holm family per selector in the earlier family and one 150-cell family per selector in the crossed-grid family. The two agree on 88 of the 90 cells. Both cells on which the two disagree are for spectral deconfounding at N=80, at H=3 and H=4, and both move toward undecided rather than reversing a verdict. Only the first moves a threshold, from $N^\star$=80 to $N^\star$=120 at H=3. One further difference is a property of the correction rather than of the data. To isolate that effect we reclassify the crossed-grid T=600 runs under both a 30-cell and a 150-cell Holm family, holding the run-level values fixed. Those same runs classify spectral deconfounding at N=60 and H=1 as recovered inside the 30-cell family and as undecided inside the five times wider 150-cell family, at an unchanged mean FDP of 0.087, and that single cell is why $N^\star$ at H=1 and T=600 is 60 under the earlier family and 80 in Table 15. Neither number is wrong, and a threshold read across the two must name the correction it belongs to.

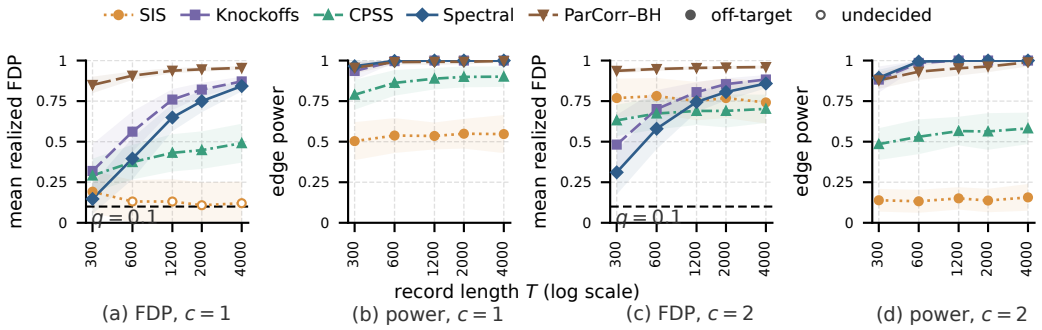


Figure 8: Record-length trajectories at N=20 and H=1, with (a,c) mean realized FDP and (b,d) edge power at c=1 and c=2, and one-SD bands over 30 runs. Filled markers denote off-target cells and hollow markers undecided cells.

D.3 RQ3: Recovery regimes and boundaries

Number of latent drivers. Thresholds across driver counts are given for the crossed-grid family in Table 15 and for the earlier (N, H) family in Table 16. The outward shift of $N^\star$ as H increases is qualitatively consistent with pervasive-confounding theory (Ćevid et al., 2020), but the benchmark uses autoregressive designs in a different dimensional regime, so it is an empirical pattern rather than an extension of that guarantee.

Table 14: Change in mean realized FDP from T=300 to T=4000 at N=20, H=1, and c=2 under full, half, and quarter loading, with paired 95% intervals over 60 seeds. The comparison is descriptive and carries no verdict.

Method	Full	Half	Quarter
Spectral deconfounding	$+0.523$ [0.488, 0.557]	$+0.332$ [0.300, 0.365]	$+0.244$ [0.223, 0.265]
Knockoffs	$+0.374$ [0.336, 0.411]	$+0.255$ [0.226, 0.284]	$+0.225$ [0.204, 0.246]

Table 15: $N^\star(H, T)$ of the three classified selectors on the crossed grid at c=2, with rows collapsed where thresholds coincide across H. A threshold is right-censored at >120 only when every evaluated cross-section is classified off-target, and unresolved when the frontier cell is undecided. Withheld marks a recovered cell followed by a non-recovered one, so no scalar threshold is reported.

Selector	H	T=300	600	1200	2000	4000
Spectral deconfounding	1	40	80	120	120	>120
	2	60	80	120	>120	>120
	3	60	120	unres.	>120	>120
	4	60	120	unres.	>120	>120
	5	80	120	unres.	>120	>120
Knockoffs screen	1–5	withheld	unres.	>120	>120	>120
CPSS	1–5	>120	>120	>120	>120	>120

Selector dependence. Table 16 is the cell-level authority for the earlier (N, H) family at T=600, whose H=1 threshold is the headline row of Table 3. The classified block contains the prespecified three-selector family; SIS and ParCorr–BH are descriptive additions and receive neither a verdict nor a threshold. Power is printed with FDP because low FDP obtained by selecting little is not equivalent to useful recovery.

Boundary sensitivity. Table 17 gives the cell-level evidence behind the sensitivity rows of Table 3. The headline loading is omitted from Table 17 because it repeats Table 16. The Holm families are one per selector over the loading-by-N cells (18, including the omitted headline loading), one per selector over the lag-by-N cells (12), and one per selector and level over the (N, H) grid (30, of which the H=1 rows are shown). The recovered Knockoff cell at $L = 3, N = 120$ has only moderate power (0.567), so this isolated recovery should not be read as evidence that enlarging the candidate set improves calibration.

Edge prevalence. Because the candidate universe grows as N^2L and systems at different N are generated separately, a change in edge prevalence could in principle alter selection difficulty without any change in estimability. Instantiating the generator alone at the headline configuration, with no selector run, over 10 runs, the candidate set and the direct-edge truth grow together, so the prevalence $|E^\star|/|\mathcal{C}|$ stays within 0.030 to 0.033 from N=10 to N=120 against a roughly twentyfold fall in Spectral realized FDP over the same range. This control is descriptive and receives no verdict. It bounds one explanation and no other, since the layout, the absolute edge count, the degree distribution, and the loading breadth all change with N as well.

D.4 Post-hoc knockoff diagnostic

We additionally evaluate the TSKI knockoff screen (Chi et al., 2025) under a post-hoc diagnostic design with N=20, H=1, T=600, L=2, and $c \in \{0, 2\}$ over five generator instances at each confounding level (Table 18). These runs lie outside the prespecified recovery-classification families, so we report descriptive statistics only and assign no recovery verdict. Mean FDP already exceeds the reference level at c=0 (0.23) despite near-perfect power, and rises to 0.80 at c=2 while power remains high. LPCMCI receives no edge-level FDP or power. Its native output is a partial ancestral graph, whose directed mark asserts ancestral direction together with adjacency rather than a one-hop direct edge, while the benchmark truth $E^\star$ is the support of $\mathbf{\Theta}$ and therefore one-hop. The benchmark defines no latent-projected or ancestral ground truth to score that output against, so no direct-edge subset of it is currently defensible, which is the rule already stated in Appendix B.3.

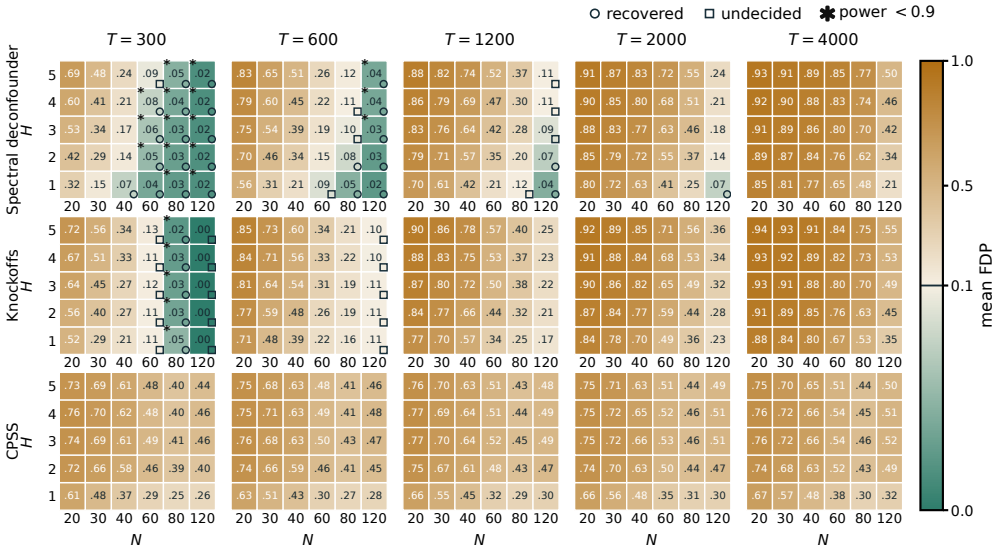


Figure 9: Mean realized FDP and frozen classification on the crossed grid at c=2, with selectors as rows, record lengths as columns, and the (N, H) grid in each panel. Symbols mark recovered and undecided cells, unmarked cells are off-target, and a star marks a recovered cell whose mean edge power is below 0.9.

Table 16: Mean realized FDP / edge power over 30 runs for the earlier (N, H) family at c=2 and T=600. Fills give the classifier verdict (green recovered, orange off-target, grey undecided), and the last column gives $N^{\star}_{M}(H)$ for the classified selectors only. The descriptive rows are not classified.

Selector	H	N=20	N=30	N=40	N=60	N=80	N=120	$N^{\star}_{M}(H)$
Pre-specified classification family								
Knockoff	1	0.70/0.99	0.51/0.98	0.36/0.98	0.22/0.97	0.17/0.96	0.10/0.90	unresolved
	2	0.78/0.97	0.60/0.96	0.48/0.96	0.24/0.95	0.18/0.93	0.11/0.85	unresolved
	3	0.80/0.96	0.68/0.95	0.52/0.94	0.29/0.93	0.20/0.91	0.10/0.79	unresolved
	4	0.83/0.96	0.71/0.93	0.57/0.93	0.33/0.91	0.21/0.88	0.10/0.75	unresolved
	5	0.86/0.94	0.74/0.91	0.60/0.90	0.34/0.89	0.21/0.84	0.10/0.71	unresolved
Spectral	1	0.58/0.99	0.35/1.00	0.20/1.00	0.08/1.00	0.05/0.99	0.02/0.96	60
	2	0.68/0.98	0.46/0.98	0.34/0.98	0.14/0.99	0.07/0.98	0.03/0.93	80
	3	0.74/0.96	0.55/0.97	0.40/0.97	0.19/0.97	0.09/0.96	0.03/0.88	80
	4	0.78/0.95	0.62/0.95	0.46/0.96	0.22/0.97	0.12/0.94	0.03/0.84	120
	5	0.83/0.94	0.66/0.93	0.53/0.94	0.26/0.95	0.12/0.92	0.04/0.80	120
CPSS	1	0.67/0.53	0.53/0.48	0.41/0.45	0.29/0.38	0.26/0.30	0.29/0.18	>120
	2	0.76/0.45	0.67/0.38	0.59/0.34	0.45/0.31	0.42/0.25	0.47/0.14	>120
	3	0.78/0.41	0.70/0.36	0.62/0.32	0.48/0.30	0.44/0.24	0.50/0.13	>120
	4	0.78/0.40	0.70/0.37	0.62/0.33	0.49/0.30	0.43/0.24	0.51/0.13	>120
	5	0.77/0.43	0.69/0.39	0.60/0.35	0.48/0.31	0.42/0.25	0.50/0.14	>120
Additional descriptive comparisons								
SIS	1	0.78/0.13	0.73/0.16	0.71/0.17	0.67/0.20	0.68/0.20	0.77/0.14	*Not classified*
	2	0.77/0.14	0.75/0.15	0.72/0.16	0.66/0.20	0.68/0.19	0.77/0.14	*Not classified*
	3	0.76/0.14	0.72/0.17	0.70/0.17	0.65/0.21	0.67/0.20	0.75/0.16	*Not classified*
	4	0.71/0.17	0.71/0.17	0.67/0.19	0.62/0.23	0.64/0.22	0.73/0.17	*Not classified*
	5	0.70/0.18	0.68/0.19	0.63/0.21	0.60/0.24	0.61/0.24	0.71/0.18	*Not classified*
ParCorr–BH	1	0.95/0.93	0.95/0.93	0.95/0.93	0.95/0.93	0.95/0.92	0.96/0.88	*Not classified*
	2	0.95/0.92	0.95/0.92	0.95/0.92	0.95/0.93	0.95/0.92	0.96/0.88	*Not classified*
	3	0.94/0.91	0.95/0.93	0.95/0.93	0.95/0.93	0.95/0.92	0.95/0.88	*Not classified*
	4	0.94/0.92	0.94/0.93	0.95/0.93	0.95/0.93	0.95/0.92	0.95/0.89	*Not classified*
	5	0.94/0.93	0.94/0.93	0.94/0.93	0.95/0.93	0.95/0.92	0.95/0.89	*Not classified*

E More Discussion

This section supplies the criteria and the source-level evidence behind the compact coverage matrix in Table 1.

Table 17: Prespecified variations of loading breadth, maximum lag, and matched level $q_M{=}q$ at $H{=}1$, $c{=}2$, and $T{=}600$, over 30 runs. In the level variation the selector's native target and the benchmark reference level move together, so each cell is tested against its own level. Maximum-lag variations regenerate the benchmark under each L rather than only widening the evaluator's candidate universe. The 95% CI is descriptive and does not determine the verdict.

Selector	Setting	N	mean FDP	95% CI	verdict	power
Loading breadth (fraction of nodes reached)						
Spectral	0.25	20	0.294	$[0.267, 0.322]$	off-target	0.993
		30	0.248	$[0.225, 0.269]$	off-target	0.996
		40	0.201	$[0.187, 0.214]$	off-target	0.996
		60	0.152	$[0.138, 0.166]$	off-target	0.997
		80	0.107	$[0.097, 0.118]$	undecided	0.997
		120	0.051	$[0.045, 0.058]$	recovered	0.986
	0.5	20	0.499	$[0.469, 0.527]$	off-target	0.991
		30	0.403	$[0.380, 0.427]$	off-target	0.995
		40	0.289	$[0.264, 0.314]$	off-target	0.995
		60	0.167	$[0.150, 0.186]$	off-target	0.997
		80	0.092	$[0.082, 0.104]$	undecided	0.995
		120	0.037	$[0.034, 0.040]$	recovered	0.980
Maximum lag L						
Knockoff	1	20	0.351	$[0.256, 0.444]$	off-target	0.759
		30	0.258	$[0.201, 0.316]$	off-target	0.957
		40	0.168	$[0.135, 0.200]$	off-target	0.969
		60	0.159	$[0.133, 0.185]$	off-target	0.981
		80	0.129	$[0.113, 0.149]$	off-target	0.985
		120	0.116	$[0.104, 0.128]$	off-target	0.989
	3	20	0.640	$[0.607, 0.672]$	off-target	0.986
		30	0.472	$[0.438, 0.505]$	off-target	0.978
		40	0.331	$[0.306, 0.355]$	off-target	0.971
		60	0.211	$[0.194, 0.227]$	off-target	0.933
		80	0.140	$[0.127, 0.151]$	off-target	0.856
		120	0.009	$[0.006, 0.012]$	recovered	0.567
Spectral	1	20	0.108	$[0.071, 0.147]$	undecided	0.979
		30	0.062	$[0.043, 0.082]$	recovered	0.989
		40	0.050	$[0.033, 0.070]$	recovered	0.991
		60	0.046	$[0.036, 0.056]$	recovered	0.994
		80	0.036	$[0.027, 0.045]$	recovered	0.993
		120	0.025	$[0.020, 0.031]$	recovered	0.995
	3	20	0.541	$[0.509, 0.572]$	off-target	0.994
		30	0.339	$[0.306, 0.369]$	off-target	0.993
		40	0.202	$[0.182, 0.223]$	off-target	0.992
		60	0.082	$[0.072, 0.093]$	recovered	0.975
		80	0.049	$[0.041, 0.059]$	recovered	0.934
		120	0.021	$[0.019, 0.022]$	recovered	0.842
Matched level $q_M{=}q$						
Spectral	0.05	20	0.486	$[0.434, 0.535]$	off-target	0.990
		30	0.261	$[0.214, 0.310]$	off-target	0.990
		40	0.130	$[0.104, 0.160]$	off-target	0.991
		60	0.047	$[0.041, 0.054]$	undecided	0.994
		80	0.024	$[0.021, 0.027]$	recovered	0.987
		120	0.010	$[0.009, 0.011]$	recovered	0.947
	0.2	20	0.667	$[0.624, 0.704]$	off-target	0.995
		30	0.474	$[0.430, 0.519]$	off-target	0.998
		40	0.314	$[0.278, 0.353]$	off-target	0.998
		60	0.155	$[0.142, 0.169]$	recovered	0.997
		80	0.095	$[0.087, 0.104]$	recovered	0.995
		120	0.055	$[0.052, 0.058]$	recovered	0.977

Table 18: Post-hoc diagnostics of the TSKI knockoff screen under the design $N{=}20$, $H{=}1$, $T{=}600$, $L{=}2$, and $c \in \{0, 2\}$. These runs lie outside the prespecified classification families and receive no recovery verdict.

Method	c	Completed	Mean FDP	Mean power	Mean sel. edges
Knockoff, TSKI (SDP)	0.0	5/5	0.228	1.000	28.8
Knockoff, TSKI (SDP)	2	5/5	0.798	0.946	101.4

E.1 Benchmark comparison criteria

Table 19 states the rule under which each column of Table 1 is checked. A check does not imply that the reported comparison isolates a pure N or T effect, and source-specific couplings are recorded in

Table 20. A blank means that the dimension was not reported in the checked paper or supplement, not that the platform could never support it.

Table 19: Qualification rule for each column of Table 1.

Criterion	Qualification rule
Ground truth	Evaluated instances have a known generating structure. An expert or constructed reference graph for real records does not suffice by itself, while a mixed benchmark qualifies when a reported subset meets the rule.
Latent confounding	An experiment with an explicit hidden common cause or a corresponding confounding construction.
Graded confounding	Multiple ordered levels, such as confounding strength, latent count or proportion, or confounder-link probability or density. A binary switch or unordered scenario categories do not qualify, and the ordered levels may be discrete.
N and T variation	Reported evaluation over changes in both variable count and record length, crossed or separate. Configurability without evaluated results does not qualify, and neither prediction horizon nor sampling frequency is treated as record length.
Multi-resolution evaluation	Separate scoring at two or more structural resolutions against the corresponding truth. Multiple metrics on one adjacency, multiple output types without resolution-specific scores, and lagged truth alone do not qualify.
Error-target assessment	Explicit comparison of an error level with a prespecified target or reference level, and an assessment of attainment. Reporting FDP, FDR, or FPR values, or using an alpha level for test selection, does not qualify by itself. The comparison need not use this benchmark's tests or verdict labels.

E.2 BENCHMARK EVIDENCE

Table 20 provides the source evidence underlying Table 1.

Table 20: Evidence underlying the benchmark comparison in Table 1.

Benchmark	Truth	Confounding	N, T variation	Resolution	Metrics / target
DGP suite (Lawrence et al., 2021)	Synthetic; exact full time graph	Latent nodes 0–20	Variable graph size; supplemental T sweep	Lag-resolved links only	F1, nSHD; no prespecified target
TimeGraph (Ferdous et al., 2025)	Synthetic; full lagged graph	Paired variants with and without a confounder	Baselines fix $N{=}4$, $T{=}500$; wider settings are generator options	One lagged-graph target	TPR, SHD, false-edge fraction; no prespecified target
TCD-Arena (Stein et al., 2026)	Synthetic SCM graphs	Five ordered sufficiency-violation levels	$T \in \{250, 1000\}$; $(N, L) \in \{(5, 3), (7, 4)\}$	Window, summary and instantaneous, scored separately	nSHD, F1, AUROC; no prespecified target
CausalCompass (Yi et al., 2026)	Synthetic SCM graph	Confounder-link density $\zeta \in \{.1, \ldots, .9\}$	$N \in \{10, 15\}$; $T \in \{500, 1000\}$	Summary/window outputs; scores use 2-D adjacencies	AUROC, AUPRC, NSHD; no prespecified target
CausalDynamics (Herdeanu et al., 2025)	Synthetic ODE/SDE and climate graphs	Confounder on/off; partial vs. full observability	$N \in \{3, 5, 10\}$; $T{=}1000$ fixed	Summary adjacency only	AUROC, AUPRC, SHD; no prespecified target
CausalTime (Cheng et al., 2024b)	Realism-oriented synthetic; derived pair graph	Causal sufficiency by default; hidden-node variant in supplement	Source domains and missing-data variant; no controlled N/T comparison	Off-diagonal directed pairs	AUROC, AUPRC; no prespecified target
CausalRivers (Stein et al., 2025)	Real records; constructed river reference	Common parents removed in confounder subgraphs	Three/five-node subgraphs; full series vs. four-week windows	Directed summary links	AUROC; no prespecified target
CauseMe/C4C (Runge et al., 2020)	Synthetic, hybrid and real; known or domain references	Unobserved-driver challenge	Evaluated N/T combinations across tasks	Lag-collapsed connectivity	ROC AUC; no prespecified target
SAVAR (Tibau et al., 2022)	Synthetic mode-level teleconnections	Not a focal variation	T and spatial resolution	Graph and mode recovery	Structural recovery metrics
PCMCI tasks (Runge et al., 2019b)	Synthetic direct lagged-edge truth	Dataset-specific conditions	Dataset- and task-specific	Lagged links	Graph-recovery metrics
Industrial benchmark (Mogensen et al., 2024)	Real records; expert reference	Not experimentally varied	Not controlled	Rolled graph	Reference-graph recovery
CausalProfiler (Panayiotou et al., 2025)	User-configured synthetic truth	Configurable; no fixed evaluation	Configurable; no fixed evaluation	User-configured	Task-dependent
MIRAGEBENCH (ours)	Synthetic; exact direct lagged-edge truth	Strength c and driver count H	Crossed (N, H, T) grid; separate c and T slices	Directed pairs and lagged edges	Pair AUPR/F1, lag accuracy, FDP/power; error-target classification in prespecified families

E.3 Adjacent work outside the formal comparison

On the (N, H) and crossed (N, H, T) grids, formal recovery classification is restricted to the three prespecified selectors with native lagged selected sets against the direct-edge truth. The separate record-length family also includes ParCorr–BH, whose lagged set is formed by the benchmark's fixed multiple-testing rule (Appendix A.1).

Adjacent estimators. Factor-augmented VARs target structural analysis using a large information set (Bernanke et al., 2005), while the published VAR(1) formulation of low-rank structured-sparse VAR estimates a sparse Granger-interaction component after accounting for latent effects, rather than a common q-controlled lagged-edge set (Basu et al., 2019). PAG and related ancestral-graph outputs likewise require latent-projected truth rather than the direct lagged-edge target used here (Gerhardus & Runge, 2020; Malinsky & Spirtes, 2018; Yehezkel Rohekar et al., 2023).

External realism benchmarks. CausalTime provides realism-oriented synthetic data with a summary-graph target, and CausalRivers provides real records with a constructed river reference (Cheng et al., 2024b; Stein et al., 2025). Under the current target semantics, they could support projected pair-level external comparisons, but neither supplies the controlled c, H, N, T axes and exact lag-resolved direct-edge truth required for formal recovery analysis.

E.4 Why a linear-Gaussian generator

The generator is a controlled instrument, not a claim about how observed systems behave. Three properties follow from that choice and are what the evaluation needs. The direct-edge target is exact, because $E^\star$ is the support of $\mathbf{\Theta}$ rather than an estimate or an expert annotation, so realized FDP is measured against a known set instead of a proxy. The latent confounding term is scaled by a single scalar, so c moves the amount of confounding while the layout, $\mathbf{\Theta}$, and the direct-edge truth stay fixed, which is what makes a structure-matched comparison across strengths possible. The axes separate, so the number of observed variables, the record length, and the latent-driver count each move without changing what the benchmark is scoring. A more realistic generator would give up at least one of these, and with it the ability to attribute a change in error control to confounding rather than to a change in the target. The cost is that every result here is a statement about this regime, which is why the recovery claims are reported per selector and per evaluated grid rather than as properties of a method.

E.5 What MirageBench does not establish

Formal verdicts cover the prespecified selector families only, and every other evaluated method is compared descriptively. A right-censored threshold lies beyond the evaluated grid if it exists; an unresolved status does not locate one. The layers report separate quantities and POLE defines no composite, so a method that ranks pairs well and controls error poorly is not ordered against one that does the reverse. Structured outputs such as partial ancestral graphs are declared but not scored, because the benchmark defines no latent-projected truth for them (Appendix D.4).

E.6 Limitations and future directions

The generator changes the latent term through c and H without separately identifying the contributions of shared latent structure, marginal scale, and total latent energy. The evaluated systems are synthetic and the grids are finite, so the boundary is located only among the sampled values. Extending the analysis to broader nonlinear dynamics would test whether the same regimes survive when the direct-effect map is no longer the identity, and real observational records with independently established causal structure would test whether they survive outside a generator altogether. A theoretical explanation of the recovery boundary remains open.