PRCD-MAP: Learning How Much to Trust Imperfect Priors in Causal Discovery

Anonymous authors
Paper under double-blind review

Abstract

External causal knowledge can improve structure learning, but incorrect or unevenly reliable priors can distort the learned graph. We propose PRCD-MAP, a framework that learns bounded trust parameters to modulate sparsity and ridge penalties in structural vector autoregression. Grouped temperatures calibrate trust across prior-probability bins; an optional contextual MLP assigns edge-specific trust. Trust updates use an empirical-Bayes-inspired agreement-and-curvature surrogate. Our analysis gives conditional estimation and local perturbation bounds, separating regularization robustness from causal identification and optimization guarantees. Experiments cover synthetic temporal graphs, CausalTime, electricity consumption, nonlinear and cross-sectional models, prior corruption, scalability, and component ablations. Cached data-driven priors improve AUROC over the no-prior backbone by 0.067 and 0.089 on AQI and Medical. In a controlled structured-error suite, learned trust reaches AUROC 0.850 versus 0.759 for a no-prior ablation, and distinguishes correct from incorrect prior assertions with AUROC 0.796. Fixed trust reaches 0.863 in that suite, and grouped temperatures lead a matched architecture comparison. Adaptive trust improves recovery in some tested regimes, but its benefits depend on sample size, prior reliability, and the optimization procedure.

1 Introduction

Limited samples and heavy noise make causal structure difficult to recover from observational data (Pearl, 2000; Spirtes et al., 2001; Runge et al., 2023; Assaad et al., 2022; Zheng et al., 2018). Practitioners in energy systems, macroeconomics, and climate science often possess partial knowledge from physical models, experts, or language models. Its reliability varies across edges and sources (Constantinou et al., 2023; Kıcıman et al., 2024). Hard constraints can discard valid edges when the supplied knowledge is incorrect; a fixed soft penalty permits uncertainty but cannot adapt its strength to the observed data.

We introduce **PRCD-MAP** (Prior-Regulated Calibrated Discovery via Maximum A Posteriori), which alternates structural coefficient estimation and bounded trust updates. The prior changes both instantaneous sparsity penalties and coefficient-wise ridge penalties. Quantile-grouped temperatures provide a compact implementation; an optional MLP uses prior and coefficient neighborhood summaries for per-edge trust. The MAP interpretation applies to coefficient estimation at fixed trust. The implemented trust objective is an EB-inspired surrogate, whose agreement labels come from the current fitted graph rather than independently observed causal edges.

Our contributions are:

- **Adaptive soft-prior consumption.** We specify the coefficient objective and the distinct grouped and contextual trust updates, including their connection to and differences from a support-conditioned Laplace approximation.
- **Conditional analysis.** We establish a local comparison to the same neutral-prior backbone and state the additional curvature, signal, identification, and optimization conditions needed for stronger conclusions. Small trust limits prior influence; learning small trust is an empirical question.
- **Evaluation across data and prior regimes.** We study sample size, noise, corruption, community heterogeneity, nonlinear models, cross-sectional graphs, electricity consumption, CausalTime, and scalability. Detailed settings, ablations, and negative controls accompany the main results.

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.

- **Direct stress tests.** Matched architecture comparisons, structured prior errors, controlled real-data priors, static Bayesian comparisons, and extended optimizer traces quantify both gains and failures. Contextual trust is useful in selected settings but does not uniformly outperform grouped or fixed trust.

2 Related Work

Causal discovery includes constraint-based (PCMCI+ (Runge, 2020); surveys (Assaad et al., 2022; Gong et al., 2025)), continuous-optimization (NOTEARS (Zheng et al., 2018), DYNOTEARS (Pamfil et al., 2020), DAGMA (Bello et al., 2022), DAGPA (Zhou et al., 2025), GRAN-DAG (Lachapelle et al., 2020), DCDFG (Lopez et al., 2022)), neural and Bayesian-posterior (Rhino (Gong et al., 2023), NGC (Tank et al., 2022), NAVAR (Bussmann et al., 2021), DiBS (Lorch et al., 2021)), and functional causal methods (VARLiNGAM (Hyvärinen et al., 2010)). Performance depends on modeling assumptions, data scaling, and benchmark design (Cheng et al., 2024; Kaiser & Sipos, 2022; Reisach et al., 2021); deep causal learning covers representation, discovery, and inference with varying assumptions and uses of external knowledge (Deng et al., 2026).

Prior integration includes hard constraints that rule edges in or out (Constantinou et al., 2023; Meek, 1995) and probabilistic structures or soft penalties (Heckerman et al., 1995). Their behavior depends on both the form and reliability of the supplied knowledge. BayesDAG (Annadani et al., 2023) performs Bayesian inference over cross-sectional causal structures. Maeda & Shimizu (2025) incorporate exclusion and temporal knowledge into additive models with unobserved variables. Differentiable Structure Learning with Partial Orders (Ban et al., 2024) uses prior ordering constraints; DAGMA-DCE (Waxman et al., 2024) studies differentiable causal-strength estimation. PRCD-MAP focuses on uncertain edge probabilities and learned, potentially edge-specific influence.

Adaptive regularization vs. learnable trust. Adaptive LASSO (Zou, 2006; Tibshirani, 1996) weights coefficients using pilot estimates, whereas SCAD (Fan & Li, 2001) uses a nonconcave coefficient-dependent penalty. Both adapt regularization internally (Bühlmann et al., 2014). PRCD-MAP treats trust as a latent hyperparameter optimized with an empirical-Bayes-inspired surrogate (Robbins, 1956; Efron, 2010). The resulting penalty depends on *external* prior signals and trust estimated from data.

LLM-generated priors. LLMs can generate plausible causal graphs (Kıcıman et al., 2024; Ban et al., 2023), but their accuracy depends on semantic information, prompting, and domain. PRCD-MAP consumes these probabilities through soft regularization (Appendix R). We distinguish semantic LLM priors from priors constructed by corrupting a benchmark reference graph. The latter measure controlled prior consumption rather than independently elicited knowledge. Appendix S compares representative approaches.

Adjacent work considers interventional Bayesian discovery, robust constraints, LLM-assisted discovery, and dynamic Bayesian networks. Our experiments cover continuous, constraint-based, and Bayesian estimators; they do not exhaust modern prior-aware alternatives. Comparisons explicitly distinguish the output target and whether a baseline receives external prior information.

3 PRCD-MAP: Prior Integration with Calibrated Trust

3.1 Problem Formulation

Consider a d-variate time series $\mathbf{X} \in \mathbb{R}^{T\times d}$ (standardized to zero mean and unit variance). Under the Structural VAR (SVAR) model, the state vector $\mathbf{x}_t \in \mathbb{R}^d$ satisfies

$$\mathbf{x}_t = \widetilde{\mathbf{W}}_0^\top \mathbf{x}_t + \sum_{k=1}^{K} \mathbf{W}_k^\top \mathbf{x}_{t-k} + \boldsymbol{\epsilon}_t, \tag{1}$$

where $\widetilde{\mathbf{W}}_0 = \mathbf{W}_0 \circ (\mathbf{1} - \mathbf{I}_d)$ is the instantaneous coefficient matrix (self-loops masked), $\mathbf{W}_k \in \mathbb{R}^{d\times d}$ is the lag-k coefficient (Lütkepohl, 2005; Sims, 1980), and $\boldsymbol{\epsilon}_t$ is i.i.d. noise. The instantaneous graph must be a DAG, enforced via a differentiable constraint $h(\widetilde{\mathbf{W}}_0) = 0$ (Zheng et al., 2018; Bello et al.,

2022). No acyclicity constraint is needed for the lagged matrices, since lagged dependencies are inherently directed forward in time.

We additionally assume a prior probability matrix $\mathbf{P}_{\text{prior}} \in [0,1]^{d\times d}$ (from input–output tables, ontologies, or LLMs), with $P_{\text{prior},ij}$ encoding expert confidence in a directed dependency $i \to j$ at *any* lag (expert/LLM elicitation is lag-agnostic by default; lag-resolved priors are an immediate extension, App. Z.4). Reliability is unknown a priori. A *Neural Additive Model* (Agarwal et al., 2021) (NAM) extension replaces the linear term with per-edge MLPs f_{ij}; the prior framework is unchanged (Appendix Q).

3.2 MAP Objective with Prior Regularization

The constrained MAP objective combines four components:

$$\min_{\mathbf{W}_{0:K}} \quad \underbrace{\mathcal{L}_{\text{data}}}_{\text{Huber fit}} + \underbrace{\mathcal{R}_{\ell_1}(\widetilde{\mathbf{W}}_0, \boldsymbol{\tau}) + \sum_{k=1}^{K} \lambda_1 \|\mathbf{W}_k\|_1}_{\text{prior-modulated sparsity}} + \underbrace{\frac{\lambda_2}{2} \sum_{k=0}^{K} \big\| \sqrt{\boldsymbol{\Omega}(\boldsymbol{\tau})} \circ \mathbf{W}_k \big\|_F^2}_{\text{prior-weighted ridge}}, \tag{2}$$

$$\text{s.t.} \quad h(\widetilde{\mathbf{W}}_0) = 0. \tag{3}$$

We describe each term below.

Data fidelity: Huber loss. To reduce sensitivity to large residuals, we use the Huber loss:

$$\mathcal{L}_{\text{data}} = \frac{1}{Td} \sum_{t,j} \ell_{\delta_H}(x_{tj} - \hat{x}_{tj}), \quad \ell_{\delta_H}(r) = \begin{cases} \frac{1}{2} r^2 & |r| \le \delta_H, \\ \delta_H(|r| - \frac{1}{2}\delta_H) & |r| > \delta_H, \end{cases} \tag{4}$$

which transitions from quadratic to linear for large residuals, down-weighting outliers.

Prior-modulated ℓ_1 penalty. We assign each edge a prior-dependent ℓ_1 coefficient:

$$\mathcal{R}_{\ell_1}(\widetilde{\mathbf{W}}_0, \boldsymbol{\tau}) = \lambda_1 \sum_{i\neq j} c_{ij}(\boldsymbol{\tau})\, |\widetilde{W}_{0,ij}|, \qquad c_{ij}(\boldsymbol{\tau}) = \text{clip}\big(1.5 - \widehat{P}_{ij}(\boldsymbol{\tau}),\ c_{\min},\ c_{\max}\big), \tag{5}$$

with $(c_{\min}, c_{\max}) = (0.1, 1.5)$. Edges with high calibrated probability $\widehat{P}_{ij}$ receive reduced ℓ_1 penalty (encouraging retention); low-probability edges receive increased penalty. The lag matrices $\mathbf{W}_{1:K}$ receive a standard ℓ_1 penalty (penalty ranges and conditioning; App. Z.3).

Prior-weighted ℓ_2 regularizer. We impose an independent Gaussian prior $W_{k,ij} \sim \mathcal{N}(0,\, \sigma_0^2/\Omega_{ij}(\boldsymbol{\tau}))$ on each coefficient, where the precision mask is

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \big(\mathbf{1}_{d\times d} - \widehat{\mathbf{P}}(\boldsymbol{\tau})\big) + \delta, \tag{6}$$

with $\delta = 10^{-3}$ preventing numerical degeneracy (so $\Omega_{ij} \in [\delta, 1+\delta]$). Edges deemed plausible by the calibrated prior receive large variance (weak shrinkage); implausible edges receive small variance (strong shrinkage toward zero). The negative log-prior yields the weighted Frobenius penalty in Eq. (2).

Calibrated prior via grouped temperature scaling. Both the ℓ_1 and ℓ_2 terms depend on the calibrated prior $\widehat{\mathbf{P}}(\boldsymbol{\tau})$. We partition the off-diagonal entries of $\mathbf{P}_{\text{prior}}$ into G groups by quantiles and assign each group a temperature parameter τ_g. The calibrated prior is

$$\widehat{P}_{ij}(\boldsymbol{\tau}) = \sigma\Big(\text{logit}\,\big(\text{clip}(P_{\text{prior},ij},\, \varepsilon,\, 1-\varepsilon)\big) \cdot \tau_{g(i,j)} \Big), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function and $\varepsilon = 10^{-3}$ prevents logit divergence at the boundaries. When $\tau_g \to 0$, the argument vanishes and $\widehat{P}_{ij} \to 0.5$, making $\boldsymbol{\Omega}$ uniform and recovering standard $\ell_1 + \ell_2$ regularization (prior ignored). When $\tau_g = 1$, the calibrated prior equals the clipped input prior. When $\tau_g > 1$, the prior is sharpened beyond its original confidence.

Structure-aware trust propagation. Grouped temperature assigns the same τ_g to all edges in a quantile bin, ignoring local topology. We refine this to a *per-edge* trust that aggregates neighborhood consistency signals. For edge (i,j) with row–column neighborhood $\mathcal{N}(i,j) = \{(k,j) : k \neq i\} \cup \{(i,l) : l \neq j\}$, we extract a 6-dimensional feature vector $\mathbf{z}_{ij}$ summarizing the prior value, neighborhood statistics of $\mathbf{P}_{\text{prior}}$ and of the current weight estimate $\mathbf{W}^*$, and a prior–data agreement term (full definition in Appendix P; implemented as $\mathbf{z}_{ij} = (P_{ij}, \bar{P}_{\mathcal{N}}, \sigma_{P_{\mathcal{N}}}, |W^*_{ij}|_{\text{norm}}, \bar{W}_{\mathcal{N}}, a_{ij})$ with $a_{ij} = 4(P_{ij}-0.5)(|W^*_{ij}|_{\text{norm}}-0.5)$). A small MLP f_θ then maps these features to a per-edge temperature:

$$\tau_{ij} \;=\; \tau_{\min} + (\tau_{\max}-\tau_{\min}) \cdot \sigma\big(f_\theta(\mathbf{z}_{ij}) + b\big), \tag{8}$$

This module uses neighborhood summaries rather than graph attention. The separate edge-GAT implementation is evaluated in Appendix AA.3. Here b is a global bias initialized from the Spearman pre-calibration. Equation (8) replaces the grouped $\tau_{g(i,j)}$ in Eq. (7) (see Def. 2, Prop. 11).

3.3 Empirical Bayes Temperature Learning

The trust parameters (grouped $\boldsymbol{\tau}$ or MLP parameters θ) control how much the optimization trusts the prior. Since prior reliability is unknown a priori, we learn them using an *EB-inspired agreement-and-curvature surrogate*.

Objective. Let $\mathbf{W}^*$ denote the current MAP estimate (held fixed when updating $\boldsymbol{\tau}$). We minimize:

$$\mathcal{L}_{\text{EB}}(\boldsymbol{\tau}) = \underbrace{\mathcal{H}\big(\widetilde{\mathbf{W}}^*_0,\ \widehat{\mathbf{P}}(\boldsymbol{\tau})\big)}_{\text{agreement loss}} + \underbrace{\frac{1}{2}\sum_{k,i,j} \log H^{(k)}_{ij}(\boldsymbol{\tau})}_{\text{diagonal curvature proxy}} + \underbrace{\frac{1}{2\sigma^2_\tau}\sum_{g=1}^{G}(\tau_g - \tfrac{1}{2})^2}_{\text{regularizer}}, \tag{9}$$

where $\mathcal{H}$ is soft binary cross-entropy with normalized edge strengths $|\widetilde{W}^*_{0,ij}|/\max|\widetilde{\mathbf{W}}^*_0|$ as soft labels and $\widehat{\mathbf{P}}(\boldsymbol{\tau})$ as predictions; the agreement term uses only instantaneous coefficients, so it does not directly validate lagged prior assertions (lag-resolved variants, MLP-regularizer, and weak-data behaviour: App. M, App. Z.4, App. Z.2). The candidate-wise diagonal curvature proxy is

$$H^{(k)}_{ij}(\boldsymbol{\tau}) = \frac{\|\mathbf{x}^{(k)}_{\cdot,i}\|_2^2}{T \cdot d} + \lambda_2\, \Omega_{ij}(\boldsymbol{\tau}). \tag{10}$$

Intuition. Agreement favors the prior when it aligns with the current estimate, but this is self-consistency rather than an independent label. The curvature term has no universal direction toward small trust. The sample-split design in Appendix V does not eliminate shared-prior dependence; its numerical summary remains provisional without the matching records. The contextual update additionally includes the prior-dependent coefficient penalties and $0.01\|\theta\|_2^2$, rather than the grouped quadratic temperature regularizer. Thus the two variants do not optimize an identical objective. The proxy omits Huber residual weights, off-diagonal curvature, support integration, and trust-dependent prior normalizers; Appendix B.9 gives the precise Laplace reference.

3.4 Optimization via Augmented Lagrangian Method

We solve Eqs. (2)–(3) via a three-level procedure. **Outer:** augmented Lagrangian $L_\rho = \mathcal{J}(\mathbf{W}_{0:K}, \boldsymbol{\tau}) + \alpha\, h(\widetilde{\mathbf{W}}_0) + \frac{\rho}{2}\, h(\widetilde{\mathbf{W}}_0)^2$; updates $\alpha \leftarrow \alpha + \rho\, h$ and $\rho \leftarrow \min(\gamma\rho, \rho_{\max})$ with γ=3. **Middle:** 8 projected gradient steps on $\mathcal{L}_{\text{EB}}(\boldsymbol{\tau})$ with $\mathbf{W}$ fixed, projected onto $[\tau_{\min}, \tau_{\max}]^G$. **Inner:** Adam (Kingma & Ba, 2015) on L_ρ for up to 400 steps with cosine LR; λ_1 inflated $5\times$ during the first $\lfloor I/3 \rfloor$ outers as warm-start (single-peaked sensitivity in App. E.2; coupled to the trust update through the fitted coefficients). Pseudocode: Algorithm 1. Finite Adam/ALM iterations give approximate solutions; a small acyclicity residual does not certify global optimality or exact nonsmooth stationarity.

3.5 Theoretical Properties

Causal interpretation requires additional identification assumptions: acyclicity alone does not identify instantaneous directions. Independent non-Gaussian innovations give one identifiable SVAR

regime (Hyvärinen et al., 2010); otherwise the output is a score-and-prior-selected DAG. Appendix B states conditional estimation, temperature, support, and contextual approximation results with proofs.

Proposition 1 (Local neutral-backbone perturbation)**.** *On a fixed acyclic support, order, and sign region, let F_0 be the neutral-prior objective and $F_\tau = F_0 + \Delta_\tau$. Suppose a convex neighborhood containing w_0, w_τ has μ-strongly convex F_0, $\nabla F_0(w_0) = 0$, and $\|\nabla F_\tau(w_\tau)\|_2 \le \eta$. If $\|\nabla \Delta_\tau(w_\tau)\|_2 \le B_\tau$, then*

$$\|w_\tau - w_0\|_2 \le (B_\tau + \eta)/\mu. \tag{11}$$

For L_R-Lipschitz prediction risk, $|R(w_\tau) - R(w_0)| \le L_R(B_\tau + \eta)/\mu$.

Here $B_\tau \le \lambda_1 \|c_\tau - c_0\|_{2,S_0} + \lambda_2 \|(\Omega_\tau - \Omega_0) \circ w_\tau\|_{2,S}$. Since $|\widehat{P}_{ij} - 1/2| \le |\operatorname{logit}(P_{ij}^{\text{clip}})\tau_{ij}|/4$, small trust limits the perturbation. The result does not prove that learned trust always becomes small, or that its predictive risk is uniformly below a no-prior estimator. The neutral reference keeps the same regularization; some empirical no-prior ablations also remove ridge and must be interpreted separately.

4 Experiments

We evaluate PRCD-MAP on: (i) synthetic data with varying prior quality and sample sizes; (ii) CausalTime benchmarks and real electricity data; (iii) component ablation. Noise robustness, scalability, cross-sectional generalization, and the cached-prior pipeline are in Appendices H–R.

4.1 Experimental Setup

Synthetic data. Ground-truth instantaneous DAGs use Erdős–Rényi graphs (d=20, edge prob. 0.15, weights from $[\pm 0.3, \pm 0.8]$). Lag-1 matrices are sparse with spectral-radius control. Series of length $T \in \{50, 100, 200, 500\}$ are simulated under Gaussian, Laplace, Student-t (ν=4), and heteroscedastic noise.

Real-world data. Three CausalTime benchmarks (Cheng et al., 2024) (AQI d=36, Medical d=20, Traffic d=20) with benchmark reference graphs from their generative construction, plus a sector-level electricity consumption dataset (d=36, Appendix C).

Prior generation. A probabilistic prior $\mathbf{P}_{\text{prior}} \in [0,1]^{d \times d}$ is constructed with controllable accuracy $\text{acc} \in \{0.4, 0.6, 0.9\}$: each entry agrees with ground truth with probability acc.

Baselines and metrics. We compare against DYNOTEARS (Pamfil et al., 2020), PCMCI+ (Runge, 2020), VARLiNGAM (Hyvärinen et al., 2010), RHINO (Gong et al., 2023), and NGC (Tank et al., 2022); DyCAST (Cheng et al., 2025) is discussed in Appendix H. Continuous variants use Adam with an augmented-Lagrangian schedule (up to 35 outer and 400 inner steps, $\text{lr} = 8 \times 10^{-3}$), with experiment-specific overrides recorded in the runners. AUROC evaluates the combined graph. Best-F1 in Table 5 uses evaluation labels to select its threshold. Tables specify replicate counts; repeated prior constructions on a fixed dataset are not independent datasets.

4.2 Effects of Prior Accuracy and Sample Size

We test how learned trust responds to reliable and unreliable priors by varying prior accuracy $\text{acc} \in \{0.4, 0.6, 0.9\}$ and sample size $T \in \{50, 100, 200, 500\}$.

Main results. At $T = 500$, the ten recorded learned-versus-fixed seed pairs give $\Delta = +0.042$ at accuracy 0.4 (two-sided paired t-test, $p = 0.033$). Differences at accuracies 0.6 and 0.9 are $+0.010$ ($p = 0.411$) and -0.004 ($p = 0.244$). These unadjusted comparisons do not give an unconditional safety guarantee. At accuracy 0.9, learned trust exceeds the three-seed PCMCI+ mean by 0.096, with different supplied prior information.

Fig. 10 shows the same sample-size settings. At $T = 500$, learned trust is 0.024 below the PCMCI+ mean at accuracy 0.4 and 0.096 above it at accuracy 0.9. These are descriptive differences across different information regimes. Appendix G reports results for additional corruption types.

Table 1: AUROC (mean±std) across sample sizes and prior accuracies (d=20, ER graph, Gaussian noise). PRCD-MAP uses 10 seeds at $T \in \{50, 500\}$ and 3 at $T \in \{100, 200\}$; external baselines use 10 at $T = 50$ and 3 otherwise (Remark 3). Baselines marked † are prior-independent (invariant to acc). **PRCD-MAP** (without suffix) uses EB-learned $\boldsymbol{\tau} \in [\tau_{\min}, \tau_{\max}]$; **PRCD-MAP ($\tau$=1)** fixes $\boldsymbol{\tau}$=**1** (full trust on the unmodulated prior, no calibration), which is *not* the no-prior baseline (the exact neutral masks arise at $P = 1/2$; experimental ablations must also be checked for their ridge setting).

Prior	Method	T=50	T=100	T=200	T=500
0.4	PRCD-MAP	0.604 ± 0.065	0.657 ± 0.015	0.712 ± 0.016	0.828 ± 0.049
	PRCD-MAP ($\tau = 1$)	0.553 ± 0.060	0.591 ± 0.017	0.650 ± 0.025	0.786 ± 0.037
0.6	PRCD-MAP	0.657 ± 0.036	0.698 ± 0.013	0.758 ± 0.013	0.869 ± 0.024
	PRCD-MAP ($\tau = 1$)	0.653 ± 0.042	0.702 ± 0.044	0.771 ± 0.029	0.859 ± 0.034
0.9	PRCD-MAP	0.765 ± 0.027	0.812 ± 0.011	0.870 ± 0.010	0.948 ± 0.013
	PRCD-MAP ($\tau = 1$)	0.791 ± 0.029	0.845 ± 0.021	0.888 ± 0.006	0.952 ± 0.015
PCMCI+†		0.651 ± 0.034	0.703 ± 0.028	0.765 ± 0.025	0.851 ± 0.034
DYNOTEARS†		0.561 ± 0.037	0.585 ± 0.025	0.622 ± 0.020	0.738 ± 0.050
VARLiNGAM†		0.580 ± 0.028	0.653 ± 0.064	0.651 ± 0.041	0.756 ± 0.028

4.3 Real-World Benchmarks

4.3.1 CausalTime Benchmark

Table 2 reports AUROC on three CausalTime datasets (Cheng et al., 2024) (semi-realistic, nonlinear, low sample sizes). We separate *practical* from *headroom* comparisons: (i) **practical** pipeline "PRCD-MAP (trust)" uses a cached data-driven prior (App. R); (ii) "PRCD-MAP (no prior)" ablates the prior entirely; (iii) **headroom** row "PRCD-MAP (oracle)" uses a ground-truth-derived probabilistic prior, indicating what the framework *could* extract from a high-quality prior, and is *not* a fair head-to-head with prior-agnostic methods. We bold within practical methods only.

Table 2: AUROC on CausalTime (T=400). PRCD-MAP (trust, cached): mean±sample std over 5 cached priors × 5 seeds = 25 fits per dataset. Priors use five correlation-based constructions (App. R). Other methods use five seeds (four recorded runs for Medical BayesDAG); zero std denotes zero observed variability. The Avg. column is the arithmetic mean of dataset means, without an uncertainty interval. Traffic cached-prior results and averages depending on them are provisional. **Bold**: best practical method. BayesDAG (Annadani et al., 2023) is cross-sectional (outputs $\mathbf{W}_0$ only); the table reports each method on its native output (capability comparison). The recorded $\mathbf{W}_0$-only comparison is in App. X; 10-seed controlled-acc re-evaluation in App. D.

Method	AQI (d=36)	Medical (d=20)	Traffic (d=20)	Avg.
PRCD-MAP (trust, cached)	**.693 ± .060**	**.583 ± .042**	.613 ± .021	**.630**
PRCD-MAP (no prior)	.626 ± .000	.494 ± .000	.611 ± .000	.577
BayesDAG (Annadani et al., 2023)	.579 ± .019	.507 ± .011	.540 ± .017	.542
PCMCI+	.570 ± .000	.540 ± .000	**.615 ± .000**	.575
RHINO	.463 ± .027	.545 ± .032	.500 ± .062	.503
VARLiNGAM	.541 ± .000	.505 ± .000	.458 ± .000	.501
NGC	.530 ± .009	.535 ± .030	.496 ± .044	.520
DYNOTEARS	.503 ± .000	.512 ± .012	.504 ± .002	.507
PRCD-MAP (oracle)†	*.703±.004*	*.587±.008*	*.686±.005*	*.659*

†Headroom row: prior derived from ground-truth adjacency; not a fair comparison with prior-agnostic baselines.

Results. Cached AQI/Medical priors give gains of $+0.123/+0.043$ over PCMCI+ and $+0.067/+0.089$ over the no-prior backbone. The Traffic cached-prior summary remains provisional because its complete per-run records have not been recovered (App. Z.6). The W_0-only BayesDAG comparison aligns the output target but not the supplied prior information (App. X).

Mechanism decomposition. The aggregate trust(cached)−PCMCI+ gap is summarized by four descriptive components: (M1) soft-prior framework, (M2) EB calibration, (M3) per-edge MLP trust, (M4) prior content; per-cell estimation in App. Z.1.

Table 3: Four-way mechanism decomposition on CausalTime (T=400). Descriptive AUROC component estimates for each enabling step (mean over 5 priors × 5 seeds). The (Backbone−PCMCI+) gap plus M1–M4 sum to the trust(cached)−PCMCI+ gap within ±0.005. This allocation is provisional: it mixes protocols and its component records and projection rule are incomplete (App. Z.1).

Dataset	Backbone − PCMCI+	(M1) Framework	(M2) EB calib.	(M3) MLP	(M4) prior content
AQI	+0.056	+0.011	+0.020	+0.011	+0.025
Medical	−0.046	+0.029	+0.020	+0.019	+0.021
Traffic	−0.004	−0.003	+0.005	+0.013	−0.013
Aggregate	+0.002	+0.012	+0.015	+0.014	+0.011

The retained allocation does not identify independent contributions. Its arithmetic agreement with the total cannot substitute for matched component-ablation records.

Scalability/nonlinear. The best PRCD-MAP variant exceeds the listed continuous-estimation baselines at the tested $d \in \{20, 50, 100, 150, 200, 250, 300\}$ (App. F); end-to-end wall-clock at d=100 is 1.7 s on GPU vs. 8,806 s for PCMCI+ on CPU (a ∼5,000× GPU/CPU wall-clock ratio; App. B.7, App. R.1). On nonlinear data the prior accuracy at which PRCD-MAP exceeds PCMCI+ decreases as d grows (App. J: **0.909** vs. 0.699 at acc=1.0, d=50); NAM (App. Q) overtakes linear at $T \geq 1000$ under bad priors.

4.4 Community-Specific Prior Reliability

To investigate the contextual heterogeneity mechanism, we use BA graphs (m=2) with hub–peripheral priors and overlapping P_{prior} ranges across communities (App. O).

Table 4: Designed validation: BA graph with hub–peripheral prior heterogeneity. Trust propagation vs. per-group temperature, mean over 3 seeds. Full 10 settings in App. O. **Bold**: larger.

		AUROC		Best-F1		Δ	
Setting	Type	trust	per-group	trust	per-group	AUROC	F1
d=20, acc=(.95, .20)	NL	**.875**	.845	**.718**	.666	+.030	+.052
d=20, acc=(.90, .30)	LIN	**.895**	.885	**.798**	.759	+.010	+.039
d=30, acc=(.95, .20)	LIN	**.946**	.931	**.812**	.773	+.015	+.039
d=30, acc=(.90, .30)	NL	**.901**	.892	**.734**	.726	+.009	+.008
Mean over all 10 settings						+.017	+.033

Trust propagation improves F1 in all 10 settings (sign test p=0.002, avg. +0.033), largest under strongest heterogeneity, in this designed family. This does not establish a universal $1/G$ performance law; the matched architecture suite favors grouped trust on aggregate (App. AA.3).

4.5 Ablation Study

Table 5: Ablation study. AUROC/best-F1 on synthetic SVAR and Lorenz-96 (d=20, T=500, acc=0.6). **Bold**: best per column. Effects depend on the dataset and component; no universal ranking follows from these settings.

	Full		LagsOnly		NoWarm		NoPrior		FixedTau		NoLam		NoL1		HardMask	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Synth.	.829	.664	**.835**	**.666**	.830	.663	.820	.657	.791	.620	.790	.615	.773	.578	.673	.563
L-96	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	.994	.949	**1.0**	.998	.981	.955	.946	.849	.720	.723

HardMask differs from Full by -0.156 (Synth.) and -0.280 (L-96) in AUROC; the FixedTau→Full gap is $+0.038$. The model also supports cross-sectional estimation; the following comparison does not isolate the effect of its instantaneous head (cross-sectional $K{=}0$: $+0.34$ vs. NOTEARS at acc$=0.4$, App. T).

4.6 Controlled Priors and Direct Diagnostic Tests

Appendix AA reports full per-condition AUROC, AUPRC, F1, and structural Hamming results for controlled CausalTime and measured cross-sectional data. These priors are generated from reference graphs; they test prior consumption rather than independent elicitation. F1 selects its threshold using evaluation labels, and SHD uses that threshold, so both are retrospective diagnostics.

In 240 structured-error fits covering block errors, direction flips, omissions, and synthetic LLM-style false positives, learned trust obtains mean AUROC 0.850 versus 0.759 for a no-prior ablation that also removes ridge. Fixed trust obtains 0.863. Correct-versus-incorrect assertion trust AUROC is 0.796, showing selective attenuation without uniformly better graph recovery. In 160 matched architecture fits, grouped trust gives AUROC/F1 0.907/0.804, compared with 0.898/0.783 for the lightweight MLP and 0.903/0.788 for edge-GAT. Across 60 extended optimizer runs, 42 retain an unchanged thresholded graph over the final five iterations; mean final-ten-iteration AUROC range is 0.000889. These tests support an optional contextual module and empirical, rather than guaranteed, optimizer stability.

5 Conclusion and Limitations

PRCD-MAP couples structural estimation with soft prior-dependent regularization and adaptive trust. Its implemented surrogate differs from exact evidence, and its perturbation result requires local regularity. Controlled experiments show improved recovery with informative priors and attenuation of structured errors. Fixed or grouped trust sometimes achieves higher recovery scores than contextual trust. Limitations include reference-derived priors in the controlled studies, oracle threshold selection for structural diagnostics, incomplete coverage of modern prior-aware baselines, nonconvex optimization, and causal identification assumptions. Deployment would require reliable label-free threshold selection and evaluation with independently elicited priors.

Reproducibility Statement

The supplementary code includes model implementations, data-processing routines, experiment runners, dependency specifications, and per-run results. Appendix AC maps each experiment to its entry point. Tables with matched per-run records are recomputed from those files. Provisional summaries lacking complete records are explicitly identified in the appendix and cannot support confirmed conclusions. Proposition 1 states its local assumptions explicitly, and Appendix B.6 supplies the proof and distinguishes exact-support conditions from thresholded-graph stability.

AI Use Statement

Generative AI tools assisted with manuscript drafting and organization, literature retrieval and reference formatting, formulation and writing of the local theoretical argument and proof, interpretation of experimental results, and code for table generation and document preparation. Numerical tables were computed from experimental result files. The authors are responsible for the scientific claims, citations, code, and final content of this work.

References

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4699–4711. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf.

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. BayesDAG: Gradient-Based Posterior Inference for Causal Discovery. In *Advances in Neural Information Processing Systems*, volume 36, pp. 1738–1763. Curran Associates, Inc., 2023. doi: 10.52202/075280-0085. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/05cf28e3d3c9a179d789c55270fe6f72-Paper-Conference.pdf.

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. URL https://doi.org/10.1613/jair.1.13428.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902v1*, 2023. URL https://arxiv.org/abs/2306.16902v1.

Taiyu Ban, Lyuzhou Chen, Xiangyu Wang, Xin Wang, Derui Lyu, and Huanhuan Chen. Differentiable Structure Learning with Partial Orders. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117426–117455. Curran Associates, Inc., 2024. doi: 10.52202/079017-3728. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d51cd79a85833b022841f7a2383b32d3-Paper-Conference.pdf.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8226–8239. Curran Associates, Inc., 2022. doi: 10.52202/068431-0598. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/36e2967f87c3362e37cf988781a887ad-Paper-Conference.pdf.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014. URL https://doi.org/10.1146/annurev-statistics-022513-115545.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural Additive Vector Autoregression Models for Causal Discovery in Time Series. In *Discovery Science*, pp. 446–460. Springer, 2021. doi: 10.1007/978-3-030-88942-5_35. URL https://doi.org/10.1007/978-3-030-88942-5_35.

Yue Cheng, Bochen Lyu, Weiwei Xing, and Zhanxing Zhu. DyCAST: Learning Dynamic Causal Structure from Time Series. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3918dd28e51ce20c38c8a1c05758e4a0-Abstract-Conference.html.

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)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0c79d6ed1788653643a1ac67b6ea32a7-Abstract-Conference.html.

Anthony C. Constantinou, Zhigao Guo, and Neville K. Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434, 2023. doi: 10.1007/s10115-023-01858-x. URL https://doi.org/10.1007/s10115-023-01858-x.

Zizhen Deng, Hu Tian, Xiaolong Zheng, and Daniel Dajun Zeng. Deep Causal Learning: Representation, Discovery and Inference. *ACM Computing Surveys*, 58(2):1–36, 2026. doi: 10.1145/3762179. URL https://doi.org/10.1145/3762179.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010. URL https://assets.cambridge.org/97811076/19678/frontmatter/9781107619678_frontmatter.pdf.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273. URL https://doi.org/10.1198/016214501753382273.

Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025. doi: 10.1038/s42256-024-00964-x. URL https://www.nature.com/articles/s42256-024-00964-x.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and YongJun Xu. Causal Discovery from Temporal Data: An Overview and New Perspectives. *ACM Computing Surveys*, 57(4): 100:1–100:38, 2025. doi: 10.1145/3705297. URL https://doi.org/10.1145/3705297.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://wenbogong.ai/publication/rhino-deep-causal-temporal-relationship-learning-with-history-dependent-noise/.

David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995. URL https://doi.org/10.1007/bf00994016.

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. URL https://www.jmlr.org/papers/v11/hyvarinen10a.html.

Marcus Kaiser and Maksim Sipos. Unsuitability of NOTEARS for causal graph discovery when dealing with dimensional quantities. *Neural Processing Letters*, 54(3):1587–1595, 2022. doi: 10.1007/s11063-021-10694-5. URL https://doi.org/10.1007/s11063-021-10694-5.

Emre Kıcıman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=mqoxLkX210.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Romain Lopez, Jan-Christian Huetter, Jonathan Pritchard, and Aviv Regev. Large-Scale Differentiable Causal Discovery of Factor Graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19290–19303. Curran Associates, Inc., 2022. doi: 10.52202/068431-1402. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/7a8fa1382ea068f3f402b72081df16be-Paper-Conference.pdf.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 24111–24123, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/ca6ab34959489659f8c3776aaf1f8efd-Abstract.html.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005. URL https://doi.org/10.1007/3-540-27752-8.

Takashi Nicholas Maeda and Shohei Shimizu. Use of prior knowledge to discover causal additive models with unobserved variables and its application to time series data. *Behaviormetrika*, 52(2): 529–547, 2025. doi: 10.1007/s41237-024-00238-1. URL https://doi.org/10.1007/s41237-024-00238-1.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 403–410, 1995. URL https://arxiv.org/abs/1302.4972.

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 Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1595–1605. PMLR, PMLR, 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, first edition, 2000. URL https://assets.cambridge.org/97805217/73621/frontmatter/9780521773621_frontmatter.pdf.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. URL https://mitpress.mit.edu/9780262037310/elements-of-causal-inference/.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e987eff4a7c7b7e580d659feb6f60c1a-Paper.pdf.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 157–163. University of California Press, 1956. URL https://digicoll.lib.berkeley.edu/record/112828?ln=en.

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, pp. 1388–1397. PMLR, PMLR, 2020. URL https://proceedings.mlr.press/v124/runge20a.html.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023. URL https://doi.org/10.1038/s43017-023-00431-y.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809. URL https://www.cs.columbia.edu/~dpeer/pub/science2005.pdf.

Christopher A Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980. URL https://doi.org/10.2307/1912017.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT Press, second edition, 2001. URL https://mitpress.mit.edu/9780262527927/causation-prediction-and-search/.

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. URL https://doi.org/10.1109/tpami.2021.3065601.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Daniel Waxman, Kurt Butler, and Petar M. Djurić. Dagma-DCE: Interpretable, Non-Parametric Differentiable Causal Discovery. *IEEE Open Journal of Signal Processing*, 5:393–401, 2024. doi: 10.1109/OJSP.2024.3351593. URL https://doi.org/10.1109/OJSP.2024.3351593.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Jincheng Zhou, Mengbo Wang, Anqi He, Yumeng Zhou, Hessam Olya, Murat Kocaoglu, and Bruno Ribeiro. Differentiable Constraint-Based Causal Discovery. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 129868–129916. Curran Associates, Inc., 2025. doi: 10.52202/085713-4325. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bc688efcf83bb744e1283897bfd4afa0-Paper-Conference.pdf.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. URL https://doi.org/10.1198/016214506000000735.

A Algorithm Details

Algorithm 1 PRCD-MAP

Input: Series $\mathbf{X}$, max lag K, prior $\mathbf{P}_{\text{prior}}$, groups G, λ_1, λ_2, ALM penalty schedule $(\rho_0, \gamma, \rho_{\max})$, outer / inner / middle iteration budgets (I, J, S), DAG-residual tolerance tol, threshold ratio β_{thr}=0.1

Output: Estimated instantaneous coefficient matrix $\widetilde{\mathbf{W}}_0$, lagged matrices $\mathbf{W}_{1:K}$

Notation: $\mathcal{J}(\mathbf{W}_{0:K}, \boldsymbol{\tau})$ denotes the unconstrained MAP objective in Eq. (2) (no DAG constraint), evaluated using the current $\boldsymbol{\tau}$ to form $c_{ij}(\boldsymbol{\tau})$ and $\Omega_{ij}(\boldsymbol{\tau})$.

Partition off-diagonal edges into G groups by quantiles of $\mathbf{P}_{\text{prior}}$
Initialize $\mathbf{W}_{0:K}$ via Ridge regression warm-start; $\alpha \leftarrow 0$, $\rho \leftarrow \rho_0$
Pre-calibrate $\boldsymbol{\tau}$ from Spearman correlation between $|\text{Corr}(\mathbf{X})|$ and $\mathbf{P}_{\text{prior}}$
for $i = 1$ **to** I **do**
    Set $\lambda_1^{(i)} \leftarrow 5\lambda_1$ if $i \le \lfloor I/3 \rfloor$, else λ_1 ▷ Lambda scheduling
    for $j = 1$ **to** J **do** ▷ Inner loop: optimize $\mathbf{W}$
        $\widetilde{\mathbf{W}}_0 \leftarrow \mathbf{W}_0 \circ (\mathbf{1} - \mathbf{I}_d)$
        Compute $L_\rho = \mathcal{J}(\mathbf{W}_{0:K}, \boldsymbol{\tau}) + \alpha\, h(\widetilde{\mathbf{W}}_0) + \frac{\rho}{2}\, h(\widetilde{\mathbf{W}}_0)^2$; update $\mathbf{W}_{0:K}$ via Adam with cosine LR
    end for
    for $s = 1$ **to** S **do** ▷ Middle level: EB update of trust parameters
        Update $\boldsymbol{\tau}$ (grouped) or θ (trust-propagation MLP) via $\nabla \mathcal{L}_{\text{EB}}$
    end for
    $h^{(i)} \leftarrow h(\widetilde{\mathbf{W}}_0)$; **if** $|h^{(i)}| < \text{tol}$ **then break** ▷ tol distinct from clip ε and noise $\boldsymbol{\epsilon}_t$
    $\alpha \leftarrow \alpha + \rho\, h^{(i)}$; $\rho \leftarrow \min(\gamma\rho,\ \rho_{\max})$
end for
Compute threshold $\theta_{\text{thr}} \leftarrow \beta_{\text{thr}} \cdot \max |\widetilde{\mathbf{W}}_0|$ from the instantaneous magnitude; zero *all* entries (across $\widetilde{\mathbf{W}}_0$ *and* $\mathbf{W}_{1:K}$) with magnitude $< \theta_{\text{thr}}$
return $\widetilde{\mathbf{W}}_0,\ \mathbf{W}_{1:K}$

B Theoretical Analysis and Proofs

The analysis separates three questions: how regularization changes a local coefficient estimate; when an estimator can approach an identifiable population target; and whether the implemented trust update selects a favorable regularization pattern. The first question admits deterministic bounds. The second needs statistical assumptions. The third is evaluated empirically and is not implied by the first two.

B.1 Notation and Preliminaries

Vectorize the allowed coefficients as $w \in \mathbb{R}^p$, where $p = (K+1)d^2 - d$: instantaneous diagonal entries are excluded, while lagged self-effects are allowed. Let S be a fixed acyclic support with a compatible instantaneous order. In a fixed sign region, the ℓ_1 term is differentiable in the active coordinates. The ridge mask repeats Ω_{ij} over enabled lags. Write $a_r = c_{ij}$ on instantaneous coefficients and $a_r = 1$ on lagged coefficients. Thus

$$F_\tau(w) = L_T(w) + \lambda_1 \sum_r a_r(\tau)|w_r| + \frac{\lambda_2}{2} \sum_r \omega_r(\tau) w_r^2.$$

The neutral comparator uses $P_{ij} = 1/2$, so $a_r = 1$ and $\omega_r = 1/2 + \delta$. It keeps λ_1, λ_2 unchanged. A positive $\tau_{\min}$ is only approximately neutral for a nonconstant prior. A no-prior experiment that also changes λ_2 uses a different comparator.

Assumption 1 (Statistical conditions for an estimation bound)**.** An identifiable target $w^\star$ is specified in an appropriate structural model. On the region used in the estimation argument, the sample loss satisfies the restricted lower-curvature inequality

$$L_T(w^\star + v) - L_T(w^\star) - \langle \nabla L_T(w^\star), v \rangle \ge \frac{\kappa_T}{2} \|v\|_2^2,$$

for the error vector under consideration, with $\kappa_T > 0$. The score obeys an explicit bound $\|\nabla L_T(w^\star)\|_\infty \le g_T$. Probability statements require these events to hold with the claimed probability under the chosen design and noise assumptions.

Stationarity, geometric mixing, and finite fourth moments are useful background conditions but do not by themselves supply a dimension-uniform exponential concentration inequality, a lower bound on the residual-weighted Huber Hessian, or causal identification. These ingredients must be established for the selected model rather than inferred from acyclicity.

Assumption 2 (Local solution regime)**.** For local differential arguments, the compared solutions lie in the same convex active-coordinate neighborhood with a fixed acyclic support and coefficient signs. The neutral objective is strongly convex there. An inexact solution has an explicitly bounded active-coordinate stationarity residual. Neither finite ALM iteration counts nor thresholded graph stability is assumed to prove these conditions.

B.2 Supporting Lemmas

Lemma 2 (Temperature scaling)**.** *Let* $u_{ij} = \mathrm{logit}\{\mathrm{clip}(P_{ij}, \epsilon, 1-\epsilon)\}$. *Then*

$$\frac{\partial \widehat{P}_{ij}}{\partial \tau} = u_{ij}\sigma(u_{ij}\tau)[1-\sigma(u_{ij}\tau)], \qquad \left|\frac{\partial \widehat{P}_{ij}}{\partial \tau}\right| \le |u_{ij}|/4.$$

Consequently $\widehat{P}_{ij}$ *moves away from* $1/2$ *in the direction of* $P_{ij} - 1/2$ *as nonnegative trust increases, and* $\widehat{P}_{ij}(0) = 1/2$.

Proof. Differentiate the sigmoid. Its derivative is positive and at most $1/4$. Integrating between zero and τ also gives $|\widehat{P}_{ij}(\tau) - 1/2| \le |u_{ij}\tau|/4$. □

Lemma 3 (Huber regularity)**.** *The Huber loss is convex and continuously differentiable, with derivative* $\mathrm{clip}(r, -\delta_H, \delta_H)$ *and 1-Lipschitz derivative. For a fixed sample and linear predictor with design matrix* Z*, the normalized sample loss has Lipschitz gradient bounded by* $\|Z\|_{\mathrm{op}}^2/(Td)$*. Its Hessian, where defined, includes residual weights* $\mathbf{1}\{|r_{tj}| < \delta_H\}$.

Proof. The derivative formula follows from the two branches of the Huber loss; the derivatives agree at their junctions. Applying the 1-Lipschitz bound to Zw and multiplying by $Z^\top/(Td)$ proves the gradient bound. Away from the junctions, a second derivative gives $Z^\top D_r Z/(Td)$, where $0 \preceq D_r \preceq I$. □

The Hessian upper bound does not establish restricted strong convexity. Ridge contributes at least $\lambda_2\delta$ to curvature in linear coefficient coordinates, but this lower bound can vanish with regularization, and it does not convexify the full DAG-constrained problem.

Lemma 4 (A sufficient restricted-curvature condition)**.** *If the residual-weighted design satisfies* $v^\top Z^\top D_r Z v/(Td) \ge \kappa_T \|v\|_2^2$ *throughout each line segment used in the local argument, then the lower-curvature inequality in Assumption 1 holds on those segments.*

Proof. Use the integral Taylor remainder along $w^\star + tv$, $t \in [0,1]$. The lower bound on the directional second derivative integrates to $\kappa_T\|v\|_2^2/2$. Huber junctions have measure zero along a nonconstant residual segment and its gradient is absolutely continuous, so the integral identity remains valid. □

Remark 1 (Dimension and conditioning). The constant κ_T is defined for the actual normalized, residual-weighted design. Stability of a VAR does not alone give a dimension-independent lower bound, and the $1/d$ loss normalization must be retained in any rate calculation. A full-rank unweighted covariance is not sufficient to establish the Huber condition when many residuals lie outside its quadratic region.

Remark 2 (Penalty range and cone constants). Since $0 < \widehat{P}_{ij} < 1$, the implemented coefficient $c_{ij} = \mathrm{clip}(1.5 - \widehat{P}_{ij}, 0.1, 1.5)$ lies in $(0.5, 1.5)$. The lower clip at 0.1 never binds for this parametrization. Including unit lag weights, $1/2 \le a_r \le 3/2$ is valid. A weighted-Lasso cone calculation, when its score-dominance assumptions hold, can therefore use weight ratio at most 3 rather than 15. This algebra does not imply a universal relation between restricted eigenvalues on different cones.

B.3 Conditional Estimation and Support Recovery

Theorem 5 (Conditional error inequality)**.** *Let $S^\star = \mathrm{supp}(w^\star)$, $s = |S^\star|$, $a_{\min} \le a_r \le a_{\max}$, and $0 \le \omega_r \le \omega_{\max}$. Suppose the lower-curvature condition holds for $\Delta = \widehat{w} - w^\star$, the estimator satisfies $F_\tau(\widehat{w}) \le F_\tau(w^\star) + e_T$ with $e_T \ge 0$, and*

$$g_T + \lambda_2 \omega_{\max} \|w^\star\|_\infty \le \lambda_1 a_{\min}/2.$$

Then, with $C_T = \lambda_1(a_{\max} + a_{\min}/2)\sqrt{s}$,

$$\|\widehat{w} - w^\star\|_2 \le \frac{C_T + \sqrt{C_T^2 + 2\kappa_T e_T}}{\kappa_T}. \tag{12}$$

This is a conditional deterministic statement for the stated objective and comparator.

Proof. The approximate objective comparison and lower-curvature condition yield

$$\begin{aligned}\frac{\kappa_T}{2}\|\Delta\|_2^2 &\le -\langle \nabla L_T(w^\star), \Delta\rangle + \lambda_1 \sum_r a_r(|w_r^\star| - |w_r^\star + \Delta_r|) \\ &\quad - \lambda_2\langle \omega \circ w^\star, \Delta\rangle - \frac{\lambda_2}{2}\sum_r \omega_r \Delta_r^2 + e_T \\ &\le b_T\|\Delta\|_1 + \lambda_1 a_{\max}\|\Delta_{S^\star}\|_1 - \lambda_1 a_{\min}\|\Delta_{(S^\star)^c}\|_1 + e_T,\end{aligned}$$

where $b_T = g_T + \lambda_2\omega_{\max}\|w^\star\|_\infty$. Apply $b_T \le \lambda_1 a_{\min}/2$ and discard the resulting nonpositive off-support term. Cauchy–Schwarz gives $\kappa_T\|\Delta\|_2^2/2 \le C_T\|\Delta\|_2 + e_T$. Solving this quadratic proves the bound. □

Remark 3 (Rate and optimization regime). An $O_p(\sqrt{s\log p/T})$ rate follows only if score concentration permits $\lambda_1 \asymp \sqrt{\log p/T}$, κ_T is bounded below in the same loss units, ridge bias satisfies the displayed domination condition, and e_T is sufficiently small. A stationary point of the nonconvex ALM problem need not satisfy the objective comparison with $w^\star$. When weights depend on data, the curvature and score event must hold uniformly over the admissible weights/solutions, or an appropriate conditional or sample-split argument must be supplied. Holding weights fixed for one numerical update does not establish statistical independence.

Support recovery. An estimation error bound alone does not give exact support. If $\|\widehat{w} - w^\star\|_\infty \le b_T$ and a fixed threshold t_T obeys $b_T < t_T < \min_{r\in S^\star}|w_r^\star| - b_T$, then $\{r : |\widehat{w}_r| > t_T\} = S^\star$. Indeed, a zero true coefficient has magnitude at most b_T, whereas a nonzero true coefficient has estimated magnitude at least $|w_r^\star| - b_T > t_T$. Such a threshold exists when the minimum signal exceeds $2b_T$. Exact unthresholded ℓ_1 support additionally needs strict inactive KKT margins and suitable design conditions. Adam output and relative-threshold graphs are not these exact supports.

B.4 Temperature Calibration: What the Surrogate Implies

Theorem 6 (Bounded sensitivity at fixed coefficients)**.** *For fixed clipped priors and fixed coefficient labels, grouped trust has bounded derivatives on its compact interval whenever the curvature proxy is bounded away from zero and numerical clipping branches are fixed. The probability sensitivity is bounded by Lemma 2. These statements do not imply that the minimizer is at $\tau_{\min}$ for an uninformative prior, or that learned temperatures increase componentwise with an accuracy parameter.*

Proof. Write a single soft-label agreement term with logit $u\tau$ as $\ell(\tau) = \log(1 + e^{u\tau}) - yu\tau$. Its derivatives are $u[\sigma(u\tau) - y]$ and $u^2\sigma(u\tau)[1 - \sigma(u\tau)] \le u^2/4$. The log-curvature term has bounded derivatives if its positive denominator stays above a fixed lower bound. The quadratic temperature penalty has constant second derivative. Finite sums retain bounded derivatives. Clipping can introduce junctions, for which this differentiable argument is local to a fixed branch. □

Remark 4 (Neutrality, calibration, and monotonicity). If $P \equiv 1/2$, both agreement and curvature are constant in grouped trust. The grouped regularizer then selects $\tau = 1/2$ when this point is feasible, not $\tau_{\min}$. Nevertheless the calibrated prior is exactly neutral. More generally, accuracy is a parameter of a prior-generation distribution, not an argument of a fixed realized loss. Population comparative statics

would require additional cross-derivative and curvature conditions. MLP parameters couple edges and can cross activation or clipping boundaries, so scalar grouped arguments do not automatically transfer to trained networks.

B.5 Heteroscedastic Noise and Local Curvature

Assumption 3 (Heteroscedastic local design)**.** Innovation variances may differ by coordinate. Identification and conditional score/curvature assumptions are imposed for this noise model, with strictly positive innovation variances and the required tail or moment controls stated separately.

Proposition 7 (Local bounds with heterogeneous variance)**.** *Proposition 1 and Theorem 5 remain valid under heterogeneous variance whenever their respective deterministic hypotheses hold, using the corresponding local curvature and score constants.*

Proof. Their proofs use strong monotonicity or an objective comparison and bounded score, rather than equality of coordinate variances. Substitute the constants for the heterogeneous design. Establishing these constants from a particular stochastic model is a separate statistical step. □

Remark 5 (Scope). Robustness of the Huber score is distinct from robustness of causal identification. Unequal Gaussian variances, correlated innovations, latent confounders, and nonstationarity may each require different identification assumptions. Noise experiments measure recovery for the specified generators, not all such models.

B.6 Prior Perturbation and Neutral-Backbone Comparison

Proof of Proposition 1. Set $v = w_\tau - w_0$. Strong convexity implies

$$\mu\|v\|_2^2 \le \langle \nabla F_0(w_\tau) - \nabla F_0(w_0), v\rangle = \langle \nabla F_\tau(w_\tau) - \nabla \Delta_\tau(w_\tau), v\rangle.$$

By Cauchy–Schwarz the last expression is at most $(\eta + B_\tau)\|v\|_2$. Divide if $v \neq 0$; otherwise the result is immediate. Apply the risk Lipschitz condition for the second claim. □

Proposition 8 (Explicit regularization perturbation)**.** *On the fixed sign region, let S_0 be its instantaneous coordinates and let $u_{ij} = \mathrm{logit}(P_{ij}^{\mathrm{clip}})$. The coefficient-gradient perturbation satisfies*

$$B_\tau \le \lambda_1\|c_\tau - c_0\|_{2,S_0} + \lambda_2\|(\Omega_\tau - \Omega_0)\circ w_\tau\|_{2,S} \tag{13}$$

$$\le \frac{\lambda_1}{4}\|u\circ\tau\|_{2,S_0} + \frac{\lambda_2}{4}\|(u\circ\tau)\circ w_\tau\|_{2,S}, \tag{14}$$

where the second term repeats the trust mask over the active lag coordinates.

Proof. The sign-fixed ℓ_1 gradient difference is $\lambda_1(c_\tau - c_0)\circ \mathrm{sign}(w_\tau)$ on S_0. Lagged ℓ_1 weights are unchanged. The ridge gradient difference is $\lambda_2(\Omega_\tau - \Omega_0)\circ w_\tau$. Apply the triangle inequality. Both masks differ from their neutral values by at most $|\widehat{P} - 1/2|$, bounded by Lemma 2. □

Remark 6 (What shrinks with trust). This bounds the change in regularization gradients, conditional on the coefficient norm and local geometry. It is not a squared-error-to-truth bound, does not automatically scale as T^{-1}, and does not assert that the learned temperature minimizes predictive risk. Positive minimum trust leaves a generally nonzero perturbation. Near-singular local curvature can amplify even small perturbations.

B.7 Computational Complexity

For dense linear coefficients, prediction and its gradient cost $O(T(K+1)d^2)$ per update. A dense matrix-based acyclicity function generally adds cubic-in-d work. Candidate-wise diagonal curvature and prior masks require $O((K+1)d^2)$ storage/work once feature norms are available. Grouped trust uses G parameters; a contextual network adds its feature extraction and forward/backward passes. Overall cost depends on the outer, inner, and trust iteration counts, stopping tolerances, and hardware.

Reported wall times are end-to-end measurements for their stated configurations. In particular, 1.7 seconds on GPU versus 8,806 seconds on CPU is a hardware-and-implementation comparison, not an algorithmic speedup under matched compute. Dense computation at $d = 300$ does not establish favorable asymptotic scaling beyond the measured range.

B.8 Finite-Iteration Temperature Learning

Theorem 9 (Conditional projected-gradient progress)**.** *Let a grouped objective Q have L-Lipschitz gradient on a convex trust box and let $\tau^+ = \mathrm{Proj}(\tau - \alpha \nabla Q(\tau))$, with $0 < \alpha \le 1/L$. Then*

$$Q(\tau^+) \le Q(\tau) - \frac{1}{2\alpha}\|\tau^+ - \tau\|_2^2.$$

Consequently the minimum squared projected-gradient mapping over m steps is at most $2[Q(\tau_0) - \inf Q]/(m\alpha)$ for this fixed objective.

Proof. Projection optimality gives $\langle \nabla Q(\tau), \tau^+ - \tau \rangle \le -\|\tau^+ - \tau\|^2/\alpha$. Combine with the smoothness inequality, using $L/2 - 1/\alpha \le -1/(2\alpha)$. Sum the descent inequality over iterations and use the gradient mapping $(\tau - \tau^+)/\alpha$. □

This is a stationarity statement, not a global-minimum guarantee for a nonconvex objective. It assumes fixed coefficients and a suitable step size; Adam updates and changing coefficient labels require separate analysis. Strong convexity would permit stronger convergence statements, but does not follow from Lipschitz gradients.

B.9 Evidence, Risk Comparison, and Safety

Support-conditioned Laplace reference. Fix an acyclic support S, order, and sign region Θ_S. A local working integral is

$$\widetilde{m}_{T,S}(X \mid \tau) = \int_{\Theta_S} e^{-F_\tau(w_S)}\, dw_S.$$

If the local mode is interior and the active-coordinate Hessian H_S is positive definite, the formal Laplace expression is

$$-\log \widetilde{m}_{T,S} \approx F_\tau(\widehat{w}_S) + \tfrac{1}{2}\log\det H_S - \tfrac{|S|}{2}\log(2\pi).$$

The integral is over active coefficients, not trust. A normalized evidence additionally requires a specified likelihood, trust-dependent prior normalizers, and a measure over supports. A normalized average Huber loss alone does not define all those factors. Near support changes or boundary modes, this local approximation needs different treatment.

Implemented surrogate. The grouped update replaces the mode objective with coefficient agreement and a quadratic trust penalty. Its diagonal curvature includes candidate coordinates and omits the Huber residual weights and off-diagonal Hessian entries. The contextual update also includes coefficient penalty values and network regularization, but still lacks the full normalizers. Hence neither is a complete marginal likelihood. The abbreviation EB denotes the motivation, not equality to the displayed integral.

Definition 1 (Expected comparison to a neutral estimator)**.** Given the joint law of data X and prior P, a procedure is ε-safe relative to w_0 for risk R if

$$\mathbb{E}_{X,P}[R(w_{\widehat{\tau}(X,P)}) - R(w_0(X))] \le \varepsilon.$$

The data, prior, fitted coefficients, and learned trust may be dependent. This is a definition; the implemented surrogate is not by itself a certificate of this property.

Corollary 10 (Conditional expected perturbation bound)**.** *If the hypotheses of Proposition 1 hold almost surely for the compared solutions and the right side is integrable, then*

$$\mathbb{E}[R(w_{\widehat{\tau}}) - R(w_0)] \le \mathbb{E}\left[\frac{L_R(B_{\widehat{\tau}} + \eta)}{\mu}\right]. \tag{15}$$

No factorization of dependent quantities is required. The bound is conditional on local regularity and has no automatic parametric rate.

Proof. Apply Proposition 1 pointwise, then take expectations. The one-sided difference is bounded by its absolute value. □

If the hypotheses hold only on an event E and the risk difference is bounded above by M elsewhere, the bound becomes the corresponding expectation on E plus $M\Pr(E^c)$. A probability bound for E must be supplied before claiming a finite-sample safety result.

Accuracy and squared prior error. For a binary flip prior that agrees with $A^\star$ with probability a, the squared error is the flip indicator. Therefore

$$\mathbb{E}(P - A^\star)^2 = 1 - a, \qquad \mathrm{Var}(\mathbf{1}\{P = A^\star\}) = a(1-a). \tag{16}$$

These are different quantities. Completely wrong priors have maximal squared error rather than zero error. Soft-valued priors require their own expectation. Perfect prior labels do not remove sample noise, weak signals, identifiability limitations, or optimizer error.

Remark 7 (Surrogate-to-risk transfer). An oracle inequality against $\inf_\tau R(w_\tau)$ needs control of the mismatch between surrogate and predictive risk, together with statistical and optimization error. Agreement with a prior-shaped coefficient estimate is insufficient to establish that mismatch. Smoothness in trust alone likewise cannot supply an $O(T^{-1})$ transfer. We report predictive and graph metrics separately from trust discrimination.

B.10 Notation Conventions and Constant Inventory

T is the effective series length; d the number of variables; K the maximum lag; G the requested number of quantile groups; $p = (K+1)d^2 - d$ the allowed coefficient count; s a combined instantaneous-and-lagged support size. $\epsilon = 10^{-3}$ clips raw probabilities, $\delta = 10^{-3}$ is the ridge-mask offset, and δ_H is the Huber threshold. These constants have distinct roles. The score bound g_T, curvature κ_T or μ, optimization errors e_T or η, and risk Lipschitz constant L_R must be evaluated in the same objective normalization. A DAG residual and an active-coordinate gradient residual are not interchangeable.

Quantile cut points are deduplicated in the implementation. Values exactly at a cut point use the lower bucket under `torch.bucketize(..., right=False)`. Tied priors can produce unequal group sizes or unused boundary buckets. Thus a group mass of exactly $1/G$ cannot be assumed in a theoretical comparison.

B.11 Connecting Theory and Experiments

The sample-size and noise sweeps test finite-sample behavior rather than certify the statistical events in Assumption 1. Thresholded support traces diagnose the output graph rather than exact active-coordinate constancy. Trust means and correctness AUROC describe selective prior attenuation rather than guaranteed risk minimization. Comparisons to a no-prior ablation must record whether ridge, warm starts, or optimization schedules also differ.

At $P \equiv 1/2$, the prior-dependent masks are exactly neutral at every temperature; this follows algebraically and requires no convergence argument. Under nonstationarity, latent confounding, or a misspecified nonlinear model, the local algebra can still hold for a working estimator, but causal interpretation and statistical rates need separate assumptions. The nonlinear additive and cross-sectional experiments evaluate extensions of the estimator; the linear identification theorem does not apply to them.

C Electricity Consumption Case Study

We apply PRCD-MAP to regional electricity consumption records obtained with permission. The processed input contains 1,105 observations and d=36 electricity-use categories, including industrial sectors, residential use, and commercial use. The loader takes seven-observation differences and uses the first 80% of the resulting sequence for training. Location and provider identifiers are withheld for anonymous review. The processed input has an integer observation index, so calendar dates cannot be reconstructed from this file alone. No ground-truth causal graph is available; this case study is a qualitative illustration, not a test of causal identification.

Fig. 1 displays the estimated instantaneous network. Links such as Chemicals–Petrochemicals and Agriculture/Forestry/Fishery–Agriculture may reflect economic association or overlapping category definitions. These interpretations are hypotheses: neither domain plausibility nor a fitted edge establishes a causal effect.

The reported temperatures reach $\tau_{\max}$ across the five runs. This records agreement between the fitted graph and the supplied prior under the implemented objective. Because the agreement labels are

derived from the fitted graph, saturation does not establish that the prior is correct; feedback between the prior and fitted coefficients is also possible.

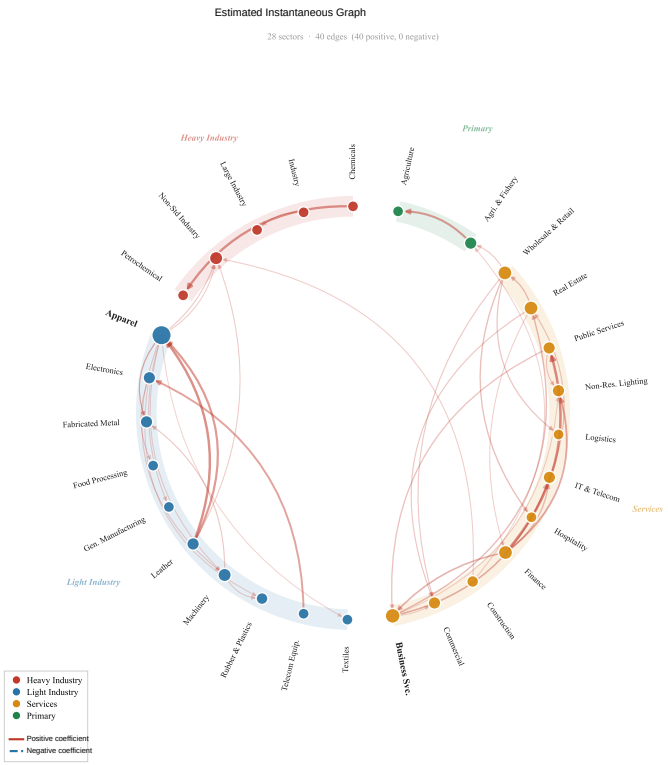


Figure 1: Estimated instantaneous network for electricity consumption (mean $|\widetilde{\mathbf{W}}_0|$ across 3 seeds). Edge weights are fitted coefficient magnitudes, not identified causal effects.

D CausalTime: 10-seed Trust Validation

We evaluate trust propagation on CausalTime with 10 seeds and controlled prior accuracies $\mathrm{acc} \in \{0.3, 0.6, 0.9\}$. This sweep compares contextual and grouped trust under reference-derived corruption. It differs from the cached-prior protocol in Table 2; its estimates must not be treated as interchangeable with the cached-prior results.

Statistical comparisons. Within each seed, we first average the trust-minus-grouped difference over the three accuracy levels. Two-sided one-sample t-tests on these ten seed averages avoid treating repeated levels of the same seed as independent replicates. The resulting dataset differences and unadjusted p-values are: AQI $\Delta = +0.031$, $p = 2.85e-05$; Traffic $\Delta = +0.013$, $p = 8.97e-04$; Medical $\Delta = +0.039$, $p = 5.67e-05$.

Per-cell comparisons. Two-sided paired t-tests use ten matched seeds in each cell. All nine comparisons are reported; p-values are unadjusted and conditional on these datasets.

- **AQI**: accuracy 0.3, $\Delta = -0.0039$, $p = 8.74e-10$; accuracy 0.6, $\Delta = -0.0089$, $p = 0.175$; accuracy 0.9, $\Delta = +0.1069$, $p = 1.25e-05$.
- **Traffic**: accuracy 0.3, $\Delta = +0.0322$, $p = 2.12e-04$; accuracy 0.6, $\Delta = -0.0002$, $p = 0.926$; accuracy 0.9, $\Delta = +0.0082$, $p = 0.061$.
- **Medical**: accuracy 0.3, $\Delta = +0.0190$, $p = 2.63e-05$; accuracy 0.6, $\Delta = -0.0012$, $p = 0.353$; accuracy 0.9, $\Delta = +0.0998$, $p = 7.30e-05$.

Table 6: CausalTime AUROC (mean±sample std over ten seeds) by prior accuracy. The avg column is the arithmetic mean over the three accuracy levels, without an uncertainty interval. Aggregate tests first average the three paired differences within each seed; baselines are prior-independent.

Dataset	Method	acc=0.3	acc=0.6	acc=0.9	avg
AQI	PRCD-MAP (trust)	0.620 ± 0.001	0.605 ± 0.014	0.817 ± 0.018	0.681
AQI	PRCD-MAP (per-group)	0.624 ± 0.001	0.614 ± 0.029	0.710 ± 0.026	0.649
AQI	DYNOTEARS	0.554 ± 0.007	0.554 ± 0.007	0.554 ± 0.007	0.554
AQI	PCMCI+	0.570 ± 0.000	0.570 ± 0.000	0.570 ± 0.000	0.570
AQI	VARLiNGAM	0.541 ± 0.000	0.541 ± 0.000	0.541 ± 0.000	0.541
Traffic	PRCD-MAP (trust)	0.589 ± 0.013	0.616 ± 0.009	0.695 ± 0.011	0.633
Traffic	PRCD-MAP (per-group)	0.557 ± 0.023	0.617 ± 0.012	0.686 ± 0.012	0.620
Traffic	DYNOTEARS	0.582 ± 0.018	0.582 ± 0.018	0.582 ± 0.018	0.582
Traffic	PCMCI+	0.615 ± 0.000	0.615 ± 0.000	0.615 ± 0.000	0.615
Traffic	VARLiNGAM	0.458 ± 0.000	0.458 ± 0.000	0.458 ± 0.000	0.458
Medical	PRCD-MAP (trust)	0.512 ± 0.004	0.545 ± 0.015	0.688 ± 0.047	0.582
Medical	PRCD-MAP (per-group)	0.493 ± 0.010	0.547 ± 0.014	0.588 ± 0.007	0.543
Medical	DYNOTEARS	0.521 ± 0.013	0.521 ± 0.013	0.521 ± 0.013	0.521
Medical	PCMCI+	0.540 ± 0.000	0.540 ± 0.000	0.540 ± 0.000	0.540
Medical	VARLiNGAM	0.505 ± 0.000	0.505 ± 0.000	0.505 ± 0.000	0.505

Four cells favor contextual trust at the unadjusted 0.05 level, one significantly favors grouped trust (AQI at accuracy 0.3), and four are nonsignificant. Nonsignificance does not establish equality. These results show a regime-dependent comparison, not uniform superiority.

E Synthetic Evaluation: 10-seed Tests and λ_1 Sensitivity

E.1 10-seed paired tests and fine accuracy grid

Table 7 retains the standard and fine-grid accuracy settings. Standard accuracies 0.4, 0.6, and 0.9 combine archived seeds 0–2 with the extension seeds 3–9. Complete per-run records for the fine-grid columns have not been located; those columns remain provisional and are excluded from confirmed significance claims.

Table 7: T=500 AUROC (mean±std) and paired-t test of *learned-τ vs. fixed-τ=1*. Cells with $\text{acc} \in \{0.4, 0.6, 0.9\}$ use 10 seeds (seeds 0–9); $\text{acc} \in \{0.3, 0.5, 0.7\}$ use seeds 0–4. PCMCI+ uses three available seeds and is prior-independent. ‡ marks provisional fine-grid summaries without matched per-run evidence. **Bold**: paired-t significant at $p<0.05$.

acc	.3	.4	.5	.6	.7	.9
Learned trust	$.844 \pm .050$	$.828 \pm .049$	$.836 \pm .051$	$.869 \pm .024$	$.872 \pm .015$	$.948 \pm .013$
Fixed trust (τ=1)	$.682 \pm .030$	$.786 \pm .037$	$.784 \pm .071$	$.859 \pm .034$	$.891 \pm .020$	$.952 \pm .015$
PCMCI+	$.851 \pm .034$	$.851 \pm .034$	$.851 \pm .034$	$.851 \pm .034$	$.851 \pm .034$	$.851 \pm .034$
Δ (learned − fixed)	$+.162^{\ddagger}$	$+\mathbf{.042}$	$+.052^{\ddagger}$	$+.010$	$-.019^{\ddagger}$	$-.004$
paired-t p-value	$.008^{\ddagger}$	$\mathbf{.033}$	$.094^{\ddagger}$	.411	$.084^{\ddagger}$	.244

Reading. The verified accuracy-0.4 difference is $+0.042$ ($p = 0.033$, unadjusted); the accuracy-0.6 and accuracy-0.9 differences are nonsignificant. The provisional fine-grid columns cannot support an average-performance or significance claim until their records are available.

E.2 λ_1 Warmup Factor: Sensitivity Sweep

Evidence status. The displayed $T = 500$, ten-seed summaries are provisional: the available sensitivity output instead contains five seeds at $T = 100$. The two protocols cannot be substituted for one another.

The retained warmup-factor table describes a proposed $T = 500$ sensitivity comparison. Matching records for that configuration remain unavailable; the existing $T = 100$ output is a different experiment. These provisional values cannot establish independence between scheduling and trust learning.

Table 8: λ_1 warmup-factor sweep at T=500, 10 seeds. AUROC (mean±std) across acc; $\bar{\tau}$ averaged over groups and seeds. All displayed values are provisional pending matching records; they are not evidence that scheduling and trust are independent.

Factor	acc=0.4 AUROC	acc=0.6 AUROC	acc=0.9 AUROC	$\bar{\tau}$ at acc=0.9
no schedule	$.750 \pm .050$	$.800 \pm .049$	$.917 \pm .030$	$1.670 \pm .130$
1×	$.750 \pm .050$	$.800 \pm .049$	$.917 \pm .030$	$1.670 \pm .130$
2×	$.821 \pm .052$	$.861 \pm .045$	$.940 \pm .018$	$1.721 \pm .055$
3×	$\mathbf{.837 \pm .034}$	$.868 \pm .037$	$\mathbf{.945 \pm .017}$	$1.744 \pm .029$
5× *(default)*	$.826 \pm .043$	$\mathbf{.874 \pm .038}$	$.942 \pm .018$	$1.746 \pm .018$
7×	$.807 \pm .050$	$.868 \pm .042$	$.939 \pm .021$	$1.722 \pm .047$
10×	$.794 \pm .059$	$.853 \pm .052$	$.935 \pm .019$	$1.707 \pm .049$

Reading. The tabulated point estimates have their largest values near factors 3–5, but the absent matching records preclude verification of uncertainty, significance, or a tuning-independence claim. Scheduling changes fitted coefficients and can therefore change trust updates.

F Main-text Scalability with Baselines

We complement Appendix I (timing) and Appendix K (PRCD-MAP-only AUROC sweep) with a head-to-head AUROC comparison at $d \in \{20, 50, 100, 150, 200, 250, 300\}$, ER graph, T=500, 3 seeds, the listed baselines included.

Table 9: Mean AUROC at $d \in \{20, 50, 100, 150, 200, 250, 300\}$, ER graph, T=500, 3 seeds. The table focuses on continuous-optimization baselines that scale to d=300; *PCMCI+ is reported separately* in Table 23 (App. R.1) because it is constraint-based with fundamentally different scaling behavior—PCMCI+ AUROC is 0.866 at d=20, 0.745 at d=50, 0.647 at d=100, and at $d \geq 150$ becomes infeasible under our 10 h/seed compute budget (PCMCI+ wall-clock is 374 s at d=50 and 8,806 s at d=100, super-linear; extrapolating gives $\gtrsim$ 24 h/seed at d=150 on our hardware). A dash denotes an unavailable result; NGC is absent from $d = 150$ onward and VARLiNGAM from $d = 250$ onward. No specific failure cause is inferred from a missing row. RHINO is excluded throughout because its training time exceeds 10h/seed already at d=50 on our test-suite GPU.

Method	d=20	d=50	d=100	d=150	d=200	d=250	d=300
PRCD-MAP (learn_tau)	.870	.751	**.602**	.608	.608	.612	.609
PRCD-MAP (fixed_tau)	**.872**	**.787**	.601	**.609**	**.609**	**.613**	**.611**
PRCD-MAP (uniform)	.859	.731	.569	.572	.571	.573	.569
DYNOTEARS	.738	.598	.581	.558	.546	.522	.521
VARLiNGAM	.756	.670	.581	.577	.571	—	—
NGC	.505	.505	.496	—	—	—	—

The best PRCD-MAP variant exceeds the listed continuous-estimation baselines at each displayed dimension. This does not extend to all methods: the separate timing suite in Table 23 gives PCMCI+ higher AUROC at $d = 20$ and $d = 100$. Protocol and hardware differences must be retained when comparing the two tables.

G Robustness to Prior Corruption Type

We test sensitivity to the prior-generation model. The main experiments use *random* corruption (each entry independently flipped relative to ground truth with probability $1-\text{acc}$). We also evaluate two structured corruption models:

- **Random**: each prior entry flipped independently (default).
- **Systematic**: corruption proportional to a fixed bias direction (e.g., over-estimating connectivity)—common when expert priors are over-confident.
- **Adversarial**: corruption targeted to maximize prior–truth mismatch given the accuracy budget.

Table 10 reports AUROC for PRCD-MAP (learn_τ) on the sample-size grid (d=20, ER graph, Gaussian noise, 5 seeds) under all three corruption modes. Differences between corruption types are uniformly ≤ 0.05 AUROC and within seed variance. The tested models give a similar regime-dependent profile (Sec. 4.2), although this comparison does not isolate the effect of empirical-Bayes calibration.

Table 10: PRCD-MAP (learn_τ) AUROC under three prior corruption types, sample-size grid (d=20, ER, Gaussian, 5 seeds; random row uses 10 seeds pooled across the recorded random-corruption runs).

acc	Corruption	T=50	T=100	T=200	T=500
0.4	Random	.607 ± .067	.657 ± .015	.712 ± .016	.834 ± .041
	Systematic	.607 ± .026	.643 ± .063	.716 ± .065	.829 ± .051
	Adversarial	.644 ± .054	.666 ± .017	.744 ± .047	.828 ± .058
0.6	Random	.653 ± .033	.698 ± .013	.758 ± .013	.869 ± .020
	Systematic	.626 ± .047	.694 ± .022	.738 ± .061	.826 ± .030
	Adversarial	.663 ± .034	.717 ± .027	.788 ± .038	.882 ± .017
0.9	Random	.765 ± .027	.812 ± .011	.870 ± .010	.951 ± .015
	Systematic	.678 ± .031	.737 ± .014	.799 ± .034	.896 ± .022
	Adversarial	.778 ± .020	.827 ± .020	.886 ± .021	.944 ± .017

Nonlinear breakdown. Table 11 extends the analysis to nonlinear data ($d \in \{10, 20\}$, 6 seeds). PRCD-MAP retains $\geq$ 0.79 AUROC at d=20 even under adversarial corruption with acc=0.2, comparable to random.

Table 11: Nonlinear AUROC under prior corruption types ($d \in \{10, 20\}$, 6 seeds, T=500).

	d=10			d=20		
acc	0.2	0.6	1.0	0.2	0.6	1.0
Systematic	.642 ± .100	.716 ± .058	.857 ± .099	.794 ± .063	.844 ± .044	.922 ± .029
Adversarial	.586 ± .061	.712 ± .052	.905 ± .054	.806 ± .073	.870 ± .027	.942 ± .019

The two corruption modes give qualitatively the same regime-dependent profile observed under random corruption (Sec. 4.2): PRCD-MAP's gain over fixed-$\boldsymbol{\tau}$ widens with acc, and the method is competitive with prior-agnostic baselines under low-accuracy priors in the tested corruption models.

H Noise Robustness

We investigate whether the choice of data-fit loss affects causal discovery under distributional mismatch. Table 12 reports AUROC under four noise distributions, averaged over prior accuracies and 10 random seeds.

PRCD-MAP exceeds the listed external baselines on average, although PCMCI+ is stronger under Gaussian noise and the no-prior variant is stronger in every displayed column, the prior-aware variant achieving average AUROC 0.907 and the no-prior variant 0.920 (vs. 0.881 for PCMCI+, 0.849 for VARLiNGAM, and 0.804 for DYNOTEARS). Under heavy-tailed noise (Laplace, Student-t), PRCD-MAP attains 0.932 and 0.936 respectively; the prior-free variant achieves higher AUROC (0.946–0.949), showing that these gains do not require the supplied prior in this setting; a matched loss ablation would be needed to attribute them to the Huber loss. DYNOTEARS, which uses a squared loss, scores 0.793 (Laplace) and 0.801 (Student-t), roughly 0.14 below the PRCD-MAP variants. We interpret this displayed difference descriptively.

VARLiNGAM also performs well under Laplace (0.913) and Student-t (0.911) noise, which is consistent with its non-Gaussian identifiability assumptions. PCMCI+ is relatively stable across noise types (range 0.863–0.894) in these runs; the partial-correlation implementation does not justify a claim of arbitrary nonlinear robustness.

We additionally evaluate DyCAST (Cheng et al., 2025), a recent Neural ODE-based method (ICLR 2025) that learns *dynamic* (time-varying) causal structures. Under our static SVAR setting, DyCAST achieves the lowest AUROC across all noise types (avg. 0.559), frequently diverging during training (NaN losses in $\sim$40% of runs). These outcomes concern our implementation and tuning in this static SVAR setting. They do not establish that the method intrinsically overfits static graphs or that structural priors are necessary.

Under heteroscedastic noise, the absolute changes from Gaussian noise are below 0.03 for the displayed methods except VARLiNGAM (+0.052). This particular variance perturbation does not characterize arbitrary temporal nonstationarity.

Table 12: AUROC (mean±std) by noise distribution (d=20, T=500, ER graph, averaged over prior accuracies and 10 seeds). **Bold**: best per column.

Method	Gaussian	Laplace	Student-t	Heterosc.	Avg.
PRCD-MAP	.870 ± .025	.932 ± .022	.936 ± .023	.890 ± .020	.907 ± .016
PRCD-MAP (no prior)	**.884 ± .042**	**.949 ± .025**	**.946 ± .028**	**.902 ± .028**	**.920 ± .021**
PCMCI+	.880 ± .025	.863 ± .055	.894 ± .040	.886 ± .029	.881 ± .032
VARLiNGAM	.759 ± .041	.913 ± .024	.911 ± .016	.811 ± .043	.849 ± .020
DYNOTEARS	.807 ± .025	.793 ± .030	.801 ± .039	.814 ± .023	.804 ± .024

I Computational Scalability

Figure 2 displays the available archived runtime measurements at $T = 500$, using three seeds per displayed method and dimension. Error bars are sample standard deviations. Table 23 describes a separate timing protocol with contextual trust. These measurements reflect implementations, stopping rules, and hardware; they do not isolate algorithmic complexity.

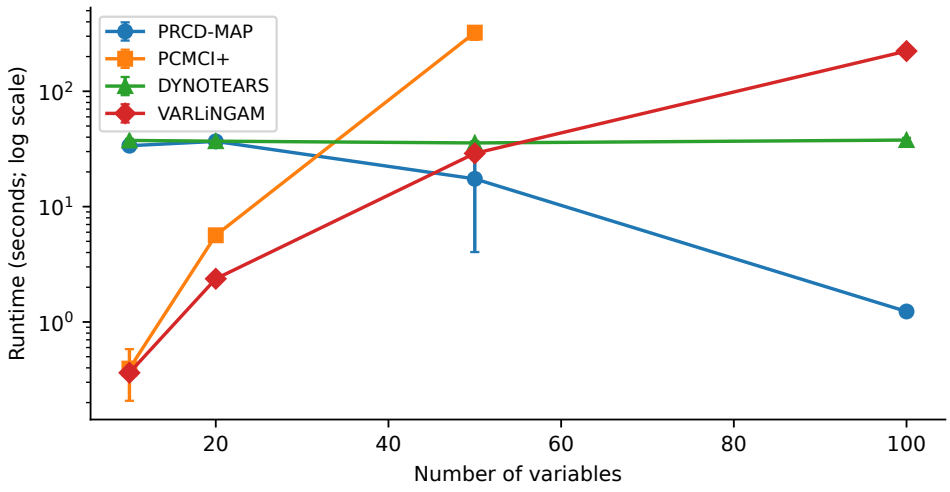


Figure 2: Available wall-clock measurements versus dimension ($T = 500$, three seeds). Points and error bars are means and sample standard deviations; the vertical axis is logarithmic. Missing measurements are not extrapolated.

J LIMITATIONS ON NONLINEAR DATA

Appendix H reports results under heavy-tailed noise within the linear SVAR model class; attributing those results specifically to the Huber loss requires a matched loss ablation. We next test nonlinear data-generating mechanisms.

Setup. We replace the linear instantaneous mechanisms in the synthetic SVAR with nonlinear functions of the form $a_{ij}\tanh(b_{ij}x_i)+c_{ij}x_i$, while keeping all other settings identical ($d \in \{10, 20\}$, T=500, ER graph, Gaussian noise, averaged over acc $\in \{0.2, 0.6, 1.0\}$ and 3 seeds). This pooled AUROC captures average performance across prior quality levels.

Results. Table 13 reports AUROC under linear vs. nonlinear instantaneous mechanisms, averaged over $d \in \{10, 20\}$ and prior-accuracy levels. PCMCI+ is largely unaffected ($0.863 \rightarrow 0.877$), despite the use of a partial-correlation test; this observation does not establish robustness to arbitrary nonlinear dependence. Among continuous-optimization methods, PRCD-MAP maintains stable AUROC ($0.747 \rightarrow 0.748$), while DYNOTEARS ($0.746 \rightarrow 0.690$) and VARLiNGAM ($0.731 \rightarrow 0.637$) degrade under model misspecification. The pooled value masks substantial regime variation: at d=20 alone, PRCD-MAP reaches 0.862 AUROC and is competitive with PCMCI+ (within 0.01); see Table 14 for the full breakdown.

Table 13: AUROC (mean±std) under linear vs. nonlinear instantaneous mechanisms (T=500, ER graph, Gaussian noise). Averaged over acc $\in \{0.2, 0.6, 1.0\}$, $d \in \{10, 20\}$, and 3 seeds.

Method	Linear	Nonlinear
PCMCI+	**.863 ± .044**	**.877 ± .047**
PRCD-MAP	.747 ± .128	.748 ± .128
DYNOTEARS	.746 ± .091	.690 ± .039
VARLiNGAM	.731 ± .044	.637 ± .106

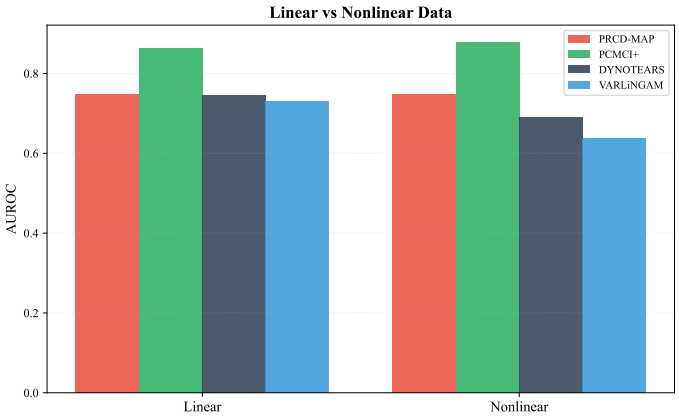


Figure 3: AUROC under linear vs. nonlinear data-generating mechanisms. PRCD-MAP is stable across regimes, while DYNOTEARS and VARLiNGAM degrade under nonlinearity.

Per-accuracy and per-dimension breakdown. The averages in Table 13 pool results across prior accuracies. Table 14 separates them by prior accuracy and dimension: the crossover between PCMCI+ and PRCD-MAP shifts to lower acc at larger d. This dependence on d is visible in the following comparisons:

Interpretation. Two possible mechanisms, not isolated by a matched ablation, may contribute to PRCD-MAP's increasing advantage at larger d: (i) the Huber loss down-weights heterogeneous residuals from nonlinear-fit-by-linear-model; (ii) the prior regularizer becomes increasingly valuable as d grows (d^2 candidate edges relative to fixed T=500 samples), so the prior signal compensates

Table 14: Nonlinear AUROC by d and prior accuracy (3 seeds, T=500; PRCD-MAP uses `learn_tau`). PCMCI+ is prior-independent in algorithmic terms; the small per-d variation across the acc rows (e.g., d=50: 0.686/0.686/0.699) reflects seed-level fluctuation in the seed pool used for each acc row, not actual prior dependence—the PCMCI+ runs at different acc rows share data but use different seed lists for paired comparison with the corresponding PRCD-MAP runs.

acc	d=20		d=30		d=50	
	PRCD-MAP	PCMCI+	PRCD-MAP	PCMCI+	PRCD-MAP	PCMCI+
0.2	.776	**.872**	.769	**.803**	.644	**.686**
0.6	.870	.872	.802	.803	**.741**	.686
1.0	**.940**	.872	**.943**	.803	**.909**	.699
avg	.862	.872	**.838**	.803	**.764**	.690

for sample scarcity. The crossover point acc* where PRCD-MAP overtakes PCMCI+ shifts from acc$^\star \approx$1.0 at d=20 to acc$^\star \approx$0.6 at d=50. PRCD-MAP has higher *average AUROC* at d=30 (+0.035) and d=50 (+0.074). The NAM extension (App. Q) provides further gains under unreliable priors at small d when sufficient samples are available ($T \geq 1000$).

K Scalability: AUROC across Dimensions

As reported in Appendix I, PRCD-MAP achieves favorable wall-clock scaling to d=100 variables. Here we complement the runtime analysis with graph recovery quality across dimensions.

Setup. We fix T=500, acc=0.6, ER graph, Gaussian noise, and vary $d \in \{10, 20, 50\}$ over 3 seeds. PCMCI+ is excluded at $d \geq 50$ due to excessive runtime (see Appendix I).

Results. Table 15 reports AUROC across $d \in \{10, 20, 50\}$. PRCD-MAP has higher AUROC at d=20 (0.829) and d=50 (0.736), outperforming DYNOTEARS and VARLiNGAM; the prior-free variant has similar scores, so these differences cannot be attributed to prior integration alone. At $d = 10$, DYNOTEARS has higher mean AUROC (0.755 versus 0.652); this comparison does not isolate the cause.

Table 15: AUROC (mean±std) across dimensions ($T = 500$, acc $= 0.6$, three seeds). The no-prior row denotes a low-trust ablation, not necessarily the same-regularization neutral estimator. PCMCI+ is absent at $d = 50$ in this sweep; separate runs are in Table 23.

Method	d=10	d=20	d=50
PRCD-MAP	.652 ± .071	**.829 ± .045**	**.736 ± .053**
PRCD-MAP (no prior)	.648 ± .068	.826 ± .042	.731 ± .049
PCMCI+	.709 ± .054	.747 ± .038	—
VARLiNGAM	.619 ± .061	.719 ± .046	.670 ± .058
DYNOTEARS	**.755 ± .042**	.738 ± .042	.599 ± .063

Recovery quality and runtime should be read separately with their own configurations. The displayed results do not establish uniform dominance across methods or dimensions.

L Optimization and Hyperparameter Analysis

This appendix supplements the ablation study (Sec. 4.5) with details on hyperparameter sensitivity and convergence behavior.

L.1 Hyperparameter Sensitivity

We sweep $\lambda_1 \in \{0.0005, 0.001, 0.003, 0.005, 0.01, 0.05\}$ and $\lambda_2 \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$: 30 configurations with three recorded seeds per configuration ($d = 20, T = 500$, acc $= 0.6$, Laplace

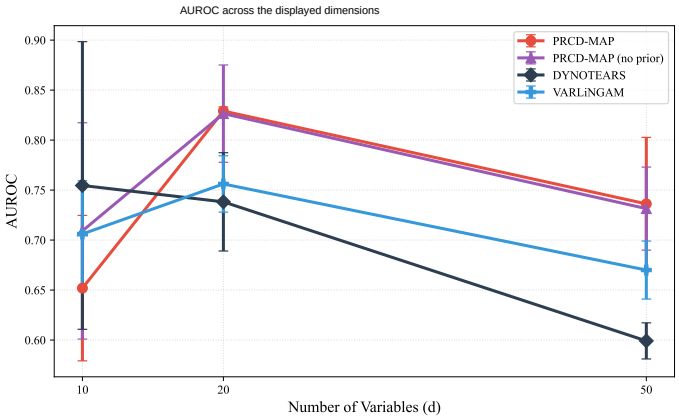


Figure 4: AUROC versus dimension for the displayed configuration ($T = 500$, acc $= 0.6$); the advantage varies with dimension and comparator.

noise). The best mean oracle-F1 is 0.830 at $\lambda_1 = 0.0005, \lambda_2 = 0.005$. For $\lambda_1 \leq 0.005$, mean F1 ranges from 0.639 to 0.830; it does not remain above 0.70 throughout. At $\lambda_1 = 0.05$, mean F1 ranges from 0.451 to 0.499.

Mean learned trust ranges from 0.050 to 1.525 across configurations. Regularization changes both fitted labels and curvature in the trust objective, so calibration is not independent of λ_1 or λ_2. Figures 5 and 6 display the recorded means.

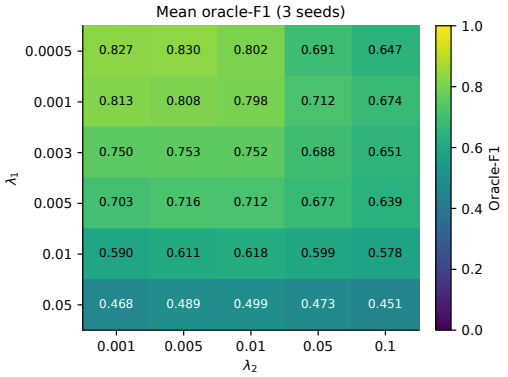


Figure 5: Mean oracle-F1 over three seeds for each (λ_1, λ_2) configuration; all 30 grid cells are shown.

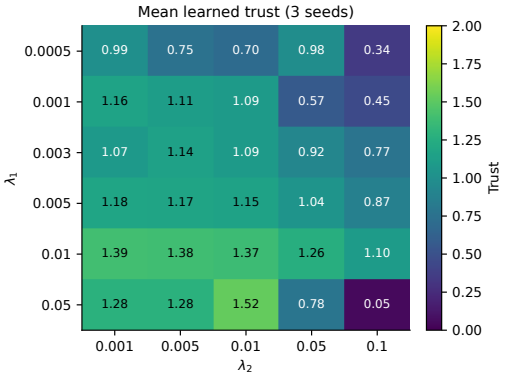


Figure 6: Mean learned trust over three seeds. The displayed variation shows dependence on regularization settings.

L.2 Convergence Analysis

Fig. 7 presents convergence diagnostics for PRCD-MAP under three prior accuracy levels (acc $\in \{0.3, 0.6, 0.9\}$), comparing the learned-$\boldsymbol{\tau}$ and fixed-$\boldsymbol{\tau}$ variants.

DAG constraint. The 18 recorded trajectories (three seeds, three accuracies, two variants) decrease monotonically in the acyclicity residual. At iteration 15, residuals range from 1.3×10^{-5} to 2.1×10^{-5}; at iteration 35 they range from 1.9×10^{-6} to 5.7×10^{-6}. These are approximate constraint violations, not exact DAG certificates.

Loss stabilization. The displayed ALM objective changes little late in the recorded runs. Its coefficients and multiplier change during training, so this trace is not a descent proof for one fixed objective. The data-fit component dominates the plotted loss values; relative component magnitudes alone do not identify their effects on graph recovery.

Role of lambda scheduling. As noted in Sec. 4.5, the `NoLam` variant of Table 5 (no λ_1 schedule, i.e., dropping the $5\times$ inflation in the first $\lfloor I/3 \rfloor$ outer iterations; equivalent to "NoLamSched" in the codebase) changes F1 by -0.049 on synthetic data. One possible explanation for the convergence traces is that the initial phase of elevated λ_1 ($5\times$ the base value for the first third of outer iterations) encourages early sparsity. Later iterations then refine the edge weights, although the traces do not establish why the schedule changes F1.

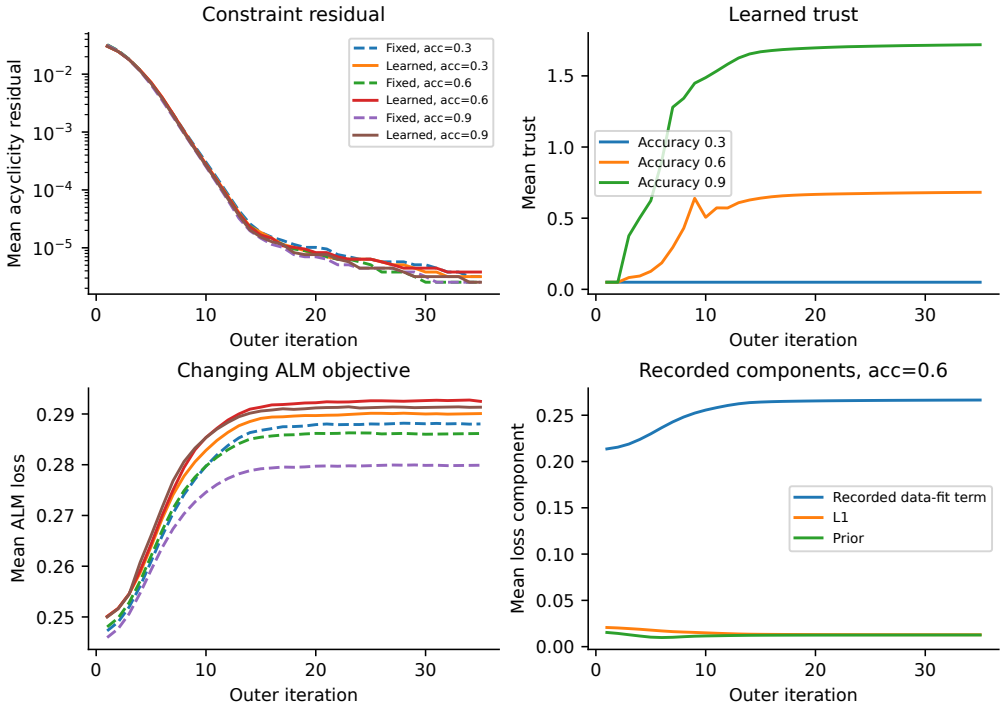


Figure 7: Convergence analysis of PRCD-MAP. Top-left: DAG constraint $h(\widetilde{\mathbf{W}}_0)$ on log scale. Top-right: $\boldsymbol{\tau}$ evolution during training. Bottom-left: total ALM loss. Bottom-right: loss decomposition at acc=0.6, showing that the data-fit component dominates.

M Temperature Learning Analysis

The learned temperature $\boldsymbol{\tau}$ controls prior influence (Sec. 4.2). We examine how $\boldsymbol{\tau}$ varies with prior quality on synthetic data.

M.1 $\boldsymbol{\tau}$ as a Function of Prior Quality

Figure 8 displays mean learned trust and oracle-F1 from the three-seed temperature sweep. Its runner uses $\tau_{\min} = 0.05$. Mean trust stays at this floor for accuracies 0–0.3, rises to 0.205, 0.618, and 1.029 at accuracies 0.4, 0.5, and 0.6, and lies between 1.629 and 1.900 at accuracies 0.7–1.0. It is not monotone: the accuracy-0.7 mean of 1.869 exceeds the accuracy-0.8 mean of 1.629.

At accuracies 0.7–1.0, learned-trust oracle-F1 ranges from 0.818 to 0.930. The uniform-prior reference has mean F1 0.903. Increasing trust is therefore not equivalent to increasing recovery quality, and fixed trust is stronger at some high-accuracy settings.

M.2 Training Dynamics of $\boldsymbol{\tau}$

Figure 9 shows one recorded trust trajectory per accuracy level. Accuracy 0.3 stays at the runner's 0.05 floor. The accuracy-0.6 and accuracy-0.9 trajectories reach 0.743 and 1.878 by iteration 35 and each contains a decrease. These traces describe finite-iteration behavior, not monotone convergence, calibrated causal probabilities, or exact acyclicity.

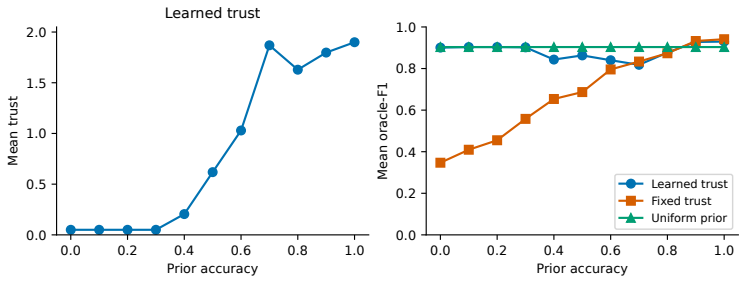


Figure 8: Mean learned trust and oracle-F1 versus prior accuracy ($d = 20, T = 500$, three seeds). Separate panels retain the different units; trust is not monotone across all levels.

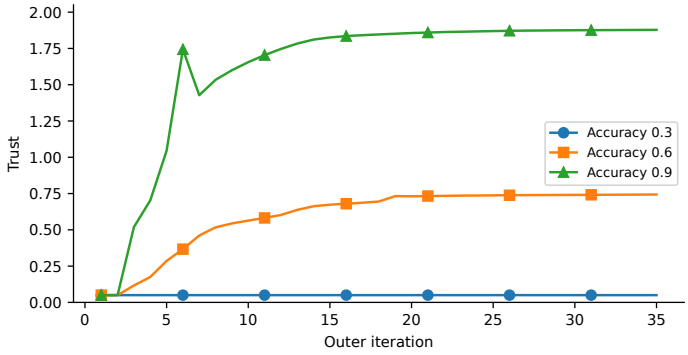


Figure 9: Trust trajectories during ALM training ($d = 20, T = 500$); each line is one recorded run. The higher-accuracy runs attain larger final temperatures in this example.

N Structure-Aware Trust Propagation: Validation

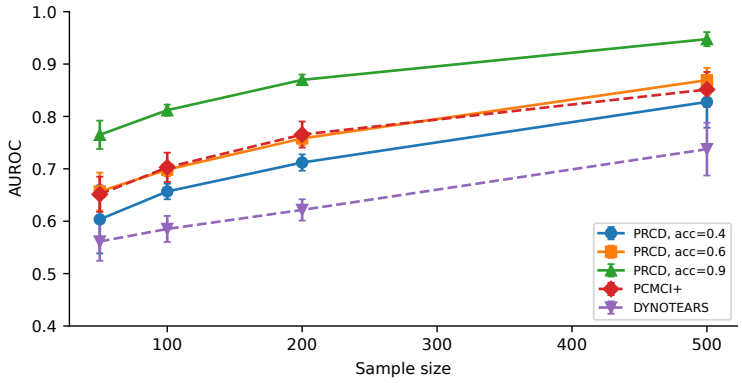


Figure 10: Prior knowledge compensates for limited data: AUROC vs. sample size T on synthetic SVAR (d=20, Gaussian noise), stratified by prior accuracy. PRCD-MAP with acc=0.9 attains mean AUROC 0.765 at $T = 50$; at acc=0.4 it remains competitive with prior-free baselines as T grows.

We validate the structure-aware trust propagation mechanism (Eq. 8, Definition 2) against the per-group temperature baseline on synthetic linear SVAR data (d=20, T=500, ER graph, Gaussian noise, 3 seeds).

Prior accuracy sweep. Table 16 reports AUROC across six prior accuracy levels under i.i.d. corruption. Differences between contextual and grouped trust are small in several settings; a 10-seed paired test at accuracy 1.0 gives an F1 gain of $+0.025$ (one-sided paired t-test $p = 0.004$, one-sided Wilcoxon $p = 0.002$), whereas the accuracy-0.8 F1 comparison has one-sided paired t-test $p = 0.287$.

Nonlinear and real-world comparisons are in Table 17. These findings describe the tested regimes; they do not turn neighborhood correlation into a quantitative safety guarantee.

Table 16: Trust propagation vs. per-group temperature on synthetic SVAR (d=20, AUROC, 3 seeds).

Method	acc=0.0	0.2	0.4	0.6	0.8	1.0
PRCD-MAP (trust)	.830	.832	.834	.849	.871	.907
PRCD-MAP (per-group)	.831	.832	.834	.848	.862	.898
PCMCI+	.851	.851	.851	.851	.851	.851
DYNOTEARS	.738	.738	.738	.738	.738	.738

Noise robustness. Across four noise types (Gaussian, Laplace, Student-t, heteroscedastic), trust propagation achieves 0.878 average AUROC vs. 0.879 for per-group. The reported Wilcoxon $p = 0.87$ is nonsignificant and does not establish equivalence or absence of degradation.

Nonlinear data. On nonlinear synthetic data ($d \in \{10, 20\}$, tanh+linear mechanisms), trust propagation provides larger gains: $+0.022$ AUROC and $+0.025$ F1 over per-group (Table 17). The improvement is concentrated at high prior accuracy ($+0.038$ F1 at d=10, acc=1.0), the highest tested prior accuracy.

Table 17: Trust propagation on nonlinear synthetic data (AUROC/F1, 3 seeds).

Method	d=10 AUROC	d=10 F1	d=20 AUROC	d=20 F1
PRCD-MAP (trust)	.843	.693	.784	.586
PRCD-MAP (per-group)	.822	.660	.760	.569
PCMCI+	.877	.714	—	—

O Community Mixing: Additional Settings and Negative Controls

We extend Sec. 4.4 with BA-graph settings at weaker heterogeneity and negative controls in which contextual trust gives mixed or negligible gains.

Additional BA settings. Table 18 reports the weaker heterogeneity setting (acc=(.85, .35)) at d=20. Improvements persist but shrink as the between-community gap narrows, suggesting sensitivity to the degree of heterogeneity in this design.

Table 18: Weaker heterogeneity on BA graphs.

Setting	AUROC		F1		Δ	
	trust	per-grp	trust	per-grp	AUROC	F1
d=20, (.85, .35), LIN	.907	.908	.812	.779	$-.001$	$+.033$
d=20, (.85, .35), NL	.851	.832	.667	.652	$+.020$	$+.015$

Negative control 1: Edge-degree heterogeneity on ER graphs. When community structure is defined by edge endpoint degrees on ER graphs (where hubs are less pronounced), improvements become inconsistent (Table 19). ER graphs have less pronounced hubs, which may make the row/column statistics less informative about community membership. This is consistent with a role for alignment between the chosen features and graph topology, but does not establish a necessary condition.

Negative control 2: Saturation under near-oracle priors. When high-accuracy communities receive near-perfect priors (acc=1.0), per-group temperature already extracts most of the signal, leaving little headroom for trust propagation (Table 20). Some mean differences over 3 seeds are negative; these comparisons do not isolate the cause. These finite-seed controls show that contextual trust does not uniformly improve recovery, including in the displayed extreme-accuracy settings.

Table 19: Negative control: ER graph with edge-degree-based heterogeneity.

Setting	trust AUROC	per-grp AUROC	trust F1	per-grp F1
d=20, (.95, .20), LIN	.870	.859	.744	.700
d=20, (.90, .30), LIN	.855	.878	.723	.744
d=20, (.85, .35), LIN	.846	.869	.711	.740

Table 20: Negative control: extreme heterogeneity on BA graphs (acc at or near oracle).

Setting	trust AUROC	per-grp AUROC	trust F1	per-grp F1
d=20, $(1.0, 0.0)$, LIN	.893	.908	.797	.821
d=20, $(1.0, 0.2)$, LIN	.912	.917	.822	.820
d=20, $(.98, 0.1)$, LIN	.917	.911	.817	.800

Summary. The designed tests show gains with graph-aligned BA heterogeneity, mixed gains and losses in the ER control, and saturation near perfect community priors. They support the contextual mechanism in these constructions. The quadratic model in Theorem 13 is illustrative; the architecture comparison in Appendix AA.3 provides a separate test in which grouped trust is strongest on aggregate.

P Contextual Trust: Information and Approximation

Definition 2 (Structure-aware trust features)**.** The contextual temperature in Eq. (8) uses $z_{ij} = (P_{ij}, \bar{P}_{\mathcal{N}}, \sigma_{P_{\mathcal{N}}}, |W_{ij}|_{\text{norm}}, \bar{W}_{\mathcal{N}}, 4(P_{ij} - 1/2)(|W_{ij}|_{\text{norm}} - 1/2))$. Its output is bounded by $\tau_{\min}$ and $\tau_{\max}$. Neighborhood summaries are computed from the available prior and fitted coefficients; they do not include true labels.

Proposition 11 (Information available in context)**.** *Let Y denote a binary edge label, P its prior value, and Z contextual features. The unrestricted optimal expected cross-entropy using P is $H(Y \mid P)$; using (P, Z) it is $H(Y \mid P, Z)$. Their difference is $I(Y; Z \mid P) \geq 0$.*

Proof. Conditional expected cross-entropy is minimized by the true conditional probability. Taking expectations gives the conditional entropies, whose difference defines conditional mutual information. □

Grouped trust still observes the individual prior through $\sigma(\tau_{g(P)} \operatorname{logit} P)$. Its optimal family risk is generally above $H(Y \mid P)$, and cannot be modeled as a predictor that only observes a bin index. A finite MLP also has approximation, estimation, and optimization errors, so the information identity does not establish its superiority.

Lemma 12 (Limits of feature-based responses)**.** *If two edges have identical six-dimensional features, a deterministic contextual network assigns them identical trust. In particular, a temperature rule requiring different outputs for such edges cannot be represented by that network.*

Proof. A function has a single output at each input. Apply this property to the identical feature vectors. □

Remark 8 (Representability and optimization). Neighborhood truth accuracy is not itself an observed feature, and correlation with binary labels is not generally an affine transformation of the fraction of correct prior assertions. Universal approximation does not reconstruct missing information or establish that a fixed small architecture realizes a target rule. It also does not guarantee that gradient optimization learns the best representable function. The matched architecture comparison therefore evaluates these choices directly (Appendix AA.3).

Q Neural Additive Model Extension

Architecture. Following the NAM framework (Agarwal et al., 2021) and its time-series extension NAVAR (Bussmann et al., 2021), each instantaneous edge $(i \to j)$ is parameterized by a 2-layer MLP $f_{ij} : \mathbb{R} \to \mathbb{R}$ with 16 hidden units. The prediction becomes $\hat{x}_{tj} = \sum_{i \neq j} f_{ij}(x_{ti}) + \sum_k \mathbf{x}_{t-k}^\top \mathbf{w}_{k,:,j}$. Edge strength $\|f_{ij}\|_{\text{param}} = (\sum_p \theta_p^2)^{1/2}$ replaces $|W_{0,ij}|$ in the DAGMA constraint and regularizers. Trust propagation uses TrustPropagationLite.

Results. We evaluate NAM across $T \in \{500, 1000, 2000\}$ on nonlinear data (d=10, 3 prior accuracies, 3 seeds). At T=500 NAM achieves 0.704 ± 0.063 AUROC, below the linear PRCD-MAP variant (0.729 at the same T). At T=1000 NAM reaches 0.780 ± 0.022, surpassing linear PRCD-MAP (0.694) by $+0.086$ AUROC. The advantage is most pronounced under unreliable priors: at acc=0.2, NAM (T=1000) attains 0.777 vs. linear 0.632, a $+0.145$ improvement. This is consistent with the per-edge MLP parameters needing more samples than T=500 provides (d=10 implies 90 MLPs); given T=1000, NAM improves recovery on nonlinear data with an unreliable prior in this comparison. NAM still trails PCMCI+ (0.892 at d=10, T=1000); this comparison does not isolate the reason for the remaining gap. At $d \geq 20$, NAM is expensive to fit (d=20→380 MLPs, ~ 30 min/run); sharing representations across edges (e.g., through a single GNN backbone) could reduce this cost.

Scalability: trust-lite at large d. On nonlinear data at $d \in \{20, 30, 50, 100\}$ (Table 21), trust-lite provides consistent improvements over per-group: $+0.018$ F1 at d=20, $+0.010$ at d=50, $+0.039$ at d=100. Runtime is $1.3-1.5\times$ that of grouped trust, an overhead of 30–50%.

Table 21: Nonlinear large-d experiments: linear model with trust propagation (F1, 3 seeds).

Method	d=20	d=30	d=50	d=100
PRCD-MAP (trust)	.678	.620	.489	.338
PRCD-MAP (per-group)	.659	.624	.479	.298

R Data-Driven Prior Construction Pipeline

We evaluate a pipeline from cached data-driven prior scores through PRCD-MAP to an estimated graph. LLM elicitation is a possible separate source of priors, but is not established by these correlation-derived CausalTime caches. Prior work studies LLM-elicited causal knowledge (Kıcıman et al., 2024; Ban et al., 2023); evaluating such priors would require a separate elicitation experiment.

Pipeline. For each CausalTime dataset, the released cache contains five data-driven matrices: absolute Pearson correlation, absolute Spearman correlation, sparse Pearson correlation retaining the top 20% of off-diagonal magnitudes (other entries set to 0.5), squared Pearson correlation, and absolute partial correlation from a covariance matrix regularized by $10^{-3}I$. These are association scores, not calibrated probabilities of directed causal edges. The same five construction rules are applied to AQI, Medical, and Traffic; their generally symmetric outputs do not supply independent directional knowledge. The cached matrices and construction metadata define the reproducible inputs. No LLM was used to generate these cached matrices.

Results. Table 22 reports AUROC averaged across the 5 priors $\times$ 5 seeds = 25 runs per (method, dataset). On AQI and Medical, the cached-prior pipeline improves AUROC over the no-prior baseline ($+0.067$ and $+0.089$ respectively); $4/5$ priors exceed the no-prior baseline on AQI and $5/5$ do so on Medical, describing variation within the selected ensemble. On Traffic, where the released variables are anonymous indices, the data-driven prior need not be neutral, and the pipeline returns $+0.002$ (within seed std); the missing Traffic records prevent confirmation of its trust distribution or significance.

Variation within the selected ensemble. The per-prior AUROC values for "trust" on AQI are $\{.645, .614, .750, .767, .688\}$, four of which exceed the no-prior baseline 0.626; on Medical,

Table 22: Cached-prior pipeline: AUROC (mean±std). Traffic entries are provisional because the matching 25-run file is unavailable. PRCD-MAP rows: 5 priors × 5 seeds = 25 runs per dataset; std is across all 25 fits; the five prior means are the units in the paired comparisons. No-Prior is deterministic given fixed data.

Method	AQI (d=36)	Medical (d=20)	Traffic (d=20)
PRCD-MAP (trust)	$\mathbf{.693 \pm .060}$	$\mathbf{.583 \pm .042}$	$\mathbf{.613 \pm .021}$
PRCD-MAP (per-group)	$.644 \pm .023$	$.517 \pm .021$	$.612 \pm .020$
No-Prior	$.626 \pm .000$	$.494 \pm .000$	$.611 \pm .000$
Δ trust vs. no-prior	$+0.067$	$+0.089$	$+0.002$
priors above no-prior	$4/5$	$5/5$	$3/5$

$\{.612, .598, .584, .608, .514\}$, $5/5$ above no-prior 0.494. This reports sensitivity within the selected ensemble; it does not rule out selection bias in constructing that ensemble. The aggregate test involving Traffic is provisional (App. Y).

Scope of the cached-prior results. The verified AQI and Medical records show positive ensemble mean differences. The Traffic cached-prior summary remains provisional. Five selected prior constructions do not establish performance for all possible priors or a population-level safety theorem. The per-prior results quantify variation within this ensemble.

R.1 Scalability Results with Trust Propagation Runtime

Table 23 extends the scalability analysis with trust propagation runtime. At d=100, trust-lite adds < 0.5 s overhead (1.7 s vs. 1.2 s for per-group), while PCMCI+ requires 8,806 s ($\sim$2.4 hours).

PRCD-MAP wall-clock measurements. The wall-clock column for PRCD-MAP variants in Table 23 is non-monotonic in d (69 s at d=10, 9 s at d=20, 34 s at d=50, 1.7 s at d=100); we extended the sweep to $d \in \{150, 200\}$ and observed the same plateau (1.7 s, 1.8 s respectively, 3 seeds each). To localize the cause we re-ran d=100 with the ALM outer-loop tolerance forced to tol=0, disabling the $|h| <$ tol break in Algorithm 1; wall-clock and AUROC were identical to the default tol=10^{-6} run across 5 seeds (mean 1.5 s, AUROC 0.611, std < 0.01 on both axes). Disabling the ALM outer-loop break had no measured effect in these runs. A possible explanation is the inner Adam early-stop (loss-stall patience of 50 steps with tolerance 10^{-6}); the OLS warm-start may produce a nearly flat inner-loop loss at $d \geq 100$; the observed early stop does not certify proximity to a local optimum, exiting well before the J=400 cap. The reported wall-clocks are reproducible from the released code under the default tolerances and should be read as *end-to-end pipeline time* with this warm-start; they do not measure per-iteration FLOPs (which scale as the derived $\mathcal{O}(J(Td^2 + d^3))$, see App. B.7). The associated d=100 AUROC of 0.617 in Table 23 is itself robust to the tol=0 forcing, showing similar measured AUROC under the two stopping settings, without certifying local optimality.

Table 23: Scalability: AUROC and runtime (seconds) with trust propagation (T=500, acc=0.6, 3 seeds).

	d=10		d=20		d=50		d=100	
Method	AUROC	Time	AUROC	Time	AUROC	Time	AUROC	Time
Trust	.926	69	.861	9	.823	34	.617	1.7
Per-grp	.933	38	.858	6	.839	23	.617	1.2
PCMCI+	.709	1	.866	8	.745	374	.647	8806
DYNOTEARS	.842	4	.814	7	.644	27	.607	36

S Methods Comparison Summary

Table 24 summarizes selected algorithmic features. Theory entries indicate that some formal analysis is available, not identical assumptions or a comparable recovery guarantee. PRCD-MAP provides conditional local analysis; the EB entry denotes a surrogate rather than exact evidence.

Table 24: Methods comparison along prior-handling, learning, and theoretical dimensions. The prior-integration column records the configurations considered here, not every capability of each method. Soft = continuous prior probabilities; Hard = binary mask; Per-edge = trust learned for each edge separately; EB = empirical-Bayes-inspired surrogate for PRCD-MAP. "Theory" indicates a formal result of some kind (including acyclicity characterizations), not necessarily statistical consistency or excess-risk control.

Method	Prior integration	Learnable trust	Bayesian view	Theory	Time series?
NOTEARS (Zheng et al., 2018)	—	—	—	✓	—
DYNOTEARS (Pamfil et al., 2020)	—	—	—	—	✓
DAGMA (Bello et al., 2022)	—	—	—	✓	—
PCMCI+ (Runge, 2020)	—	—	—	✓	✓
RHINO (Gong et al., 2023)	—	—	✓	✓	✓
NOTEARS + hard mask	Hard	—	—	—	—
Adaptive LASSO (Zou, 2006)	—	implicit[†]	—	✓	—
BayesDAG (Annadani et al., 2023)	—	—	✓	—	—
DiBS (Lorch et al., 2021)	—	—	✓	—	—
PRCD-MAP (ours)	**Soft**	**Per-edge**	**EB**	✓	✓

[†]Adaptive LASSO modulates penalty internally via pilot estimates; it does not consume an external prior signal. PRCD-MAP keeps the supplied prior fixed and calibrates its influence through learned trust hyperparameters.

T Cross-Sectional Structure Learning

To evaluate cross-sectional estimation, we instantiate PRCD-MAP with $K{=}0$ on i.i.d. data from a linear SEM $\mathbf{X} = \mathbf{X}\mathbf{W}_0 + \boldsymbol{\epsilon}$ (Peters et al., 2017), against NOTEARS (Zheng et al., 2018), NOTEARS+mask, and DAGMA (Bello et al., 2022) ($d{=}20$, Gaussian noise, 5 seeds).

Table 25: Cross-sectional structure learning ($d{=}20$, $n{=}500$, ER, Gaussian, 5 seeds).

acc	Method	AUROC	F1	SHD
0.4	NOTEARS	.456 ± .061	.362 ± .066	31.0
	NOTEARS+mask	.420 ± .120	.263 ± .139	158.8
	DAGMA	.538 ± .055	.337 ± .065	33.2
	PRCD-MAP	**.797 ± .115**	**.589 ± .231**	**23.8**
0.6	NOTEARS	.456 ± .061	.362 ± .066	31.0
	NOTEARS+mask	.653 ± .082	.485 ± .107	23.2
	DAGMA	.538 ± .055	.337 ± .065	33.2
	PRCD-MAP	**.865 ± .067**	**.707 ± .121**	**16.2**
0.9	NOTEARS	.456 ± .061	.362 ± .066	31.0
	NOTEARS+mask	.943 ± .030	**.937 ± .039**	**3.0**
	DAGMA	.538 ± .055	.337 ± .065	33.2
	PRCD-MAP	**.948 ± .047**	.886 ± .099	6.0

PRCD-MAP has higher AUROC at acc $\in \{0.4, 0.6\}$ (over DAGMA by $+0.26$ and $+0.33$ AUROC respectively); the hard mask yields more adjacency errors (SHD 159 vs. 24 at acc$=0.4$). At acc$=0.9$, the mask achieves slightly higher F1 by exploiting near-oracle information, while PRCD-MAP attains the highest AUROC. These comparisons demonstrate a cross-sectional implementation under the stated synthetic settings; they do not establish generality across domains.

Comparison with BayesDAG. Appendix X retains the instantaneous and combined-output comparisons; Appendix AA.4 gives a separate static acyclic experiment. BayesDAG uses its native configuration without the external prior, whereas PRCD-MAP receives the controlled or cached data-driven prior. These comparisons therefore evaluate different information regimes. Aligning W_0 targets does not isolate consumption of an identical prior. NOTEARS, masked NOTEARS, and DAGMA remain useful continuous-estimation comparison points.

U Grouped and Per-Edge Trust in a Quadratic Model

Theorem 13 (Within-group approximation cost)**.** *Consider a stylized objective* $R(t) = n^{-1}\sum_r[-b_r t_r + q_r t_r^2/2]$ *with* $q_r > 0$*. Suppose the individual unconstrained optima* $t_r^* = b_r/q_r$ *all lie in the allowed trust interval. For a partition into groups, let* $\bar{t}_g = (\sum_{r\in g} q_r t_r^*)/(\sum_{r\in g} q_r)$*. Then*

$$R^*_{\text{group}} - R^*_{\text{edge}} = \frac{1}{2n}\sum_g \sum_{r\in g} q_r(\bar{t}_g - t_r^*)^2. \tag{17}$$

Proof. Complete the square: $-b_r t + q_r t^2/2 = q_r(t - t_r^*)^2/2 - b_r^2/(2q_r)$. Individual minimization gives t_r^*. Group minimization gives the weighted mean $\bar{t}_g$, which is feasible because the individual optima are feasible. Subtract the objectives and sum. □

For two subsets of a mixed group with total weights Q_1, Q_2 and distinct common optima t_1^*, t_2^*, the contribution is at least $Q_1 Q_2(t_1^* - t_2^*)^2/[2n(Q_1 + Q_2)]$. This follows from the weighted variance identity, and illustrates why genuinely different optimal temperatures in one group can matter.

Remark 9 (When the gap vanishes). Different prior-error variances alone do not imply different b_r/q_r. A mixed bin can have identical optimal temperatures and zero gap. Unequal quantile-bin masses also preclude replacing group mass by $1/G$ without an assumption. If all individual constrained optima saturate at the same boundary, grouped and per-edge optima again coincide. A universal $\Omega(1/G)$ lower bound therefore does not follow from heterogeneity alone.

Remark 10 (Relation to the implemented objective). The quadratic objective is an illustrative approximation problem. Its curvature q_r is not automatically the variance of prior errors, and no uniform Taylor remainder has been established for transferring this gap to the implemented surrogate. Its risk also differs from AUROC or F1. The community-mixing experiments are empirical tests of a mechanism, not numerical verification of the displayed quadratic bound or a guarantee for the trained MLP.

V Cross-fitted EB: Sample-split Stability of $\boldsymbol{\tau}^\star$

Evidence status. The numerical summaries below are provisional because matching per-run cross-fit records are not available in the supplement.

Prior dependence of the agreement labels. Eq. (9) uses the prior-regularized MAP estimate $\mathbf{W}^*$ as soft labels for the agreement loss $\mathcal{H}$. Because $\mathbf{W}^*$ is itself shaped by $\mathbf{P}_{\text{prior}}$ through the modulated ℓ_1 and ℓ_2 penalties, $\mathcal{H}(\widetilde{\mathbf{W}}_0^*, \widehat{\mathbf{P}}(\boldsymbol{\tau}))$ may reward prior-induced patterns in $\mathbf{W}^*$ even when the prior assertions are incorrect.

Dependence in the alternating update. Holding coefficients fixed during a trust step makes its soft labels numerically constant, but does not remove their dependence on the prior or make the final iterate a stationary point of the fully differentiated bilevel objective. The diagonal curvature proxy is independent of fitted coefficients at fixed data, yet its derivative does not always favor smaller trust. Sample splitting is therefore a sensitivity experiment rather than a proof that confirmation bias is absent.

Empirical sample-split test. We implement a cross-fitted EB variant: split $\{\mathbf{x}_t\}_{t=1}^T$ *chronologically* into halves $\mathcal{D}_A = \{\mathbf{x}_t\}_{t=1}^{T/2}$ and $\mathcal{D}_B = \{\mathbf{x}_t\}_{t=T/2+1}^{T}$ (lagged rows are formed within each half; adjacent halves remain temporally dependent, and no negligible-boundary claim follows from

Assumption 1); estimate $\mathbf{W}_A^*$ from $\mathcal{D}_A$; learn $\boldsymbol{\tau}_A^\star$ by minimizing $\mathcal{L}_{\mathrm{EB}}$ using $\mathbf{W}_A^*$ as labels but evaluating the diagonal log-curvature proxy on $\mathcal{D}_B$. The agreement labels remain fitted on $\mathcal{D}_A$ using the same prior; evaluating feature norms on $\mathcal{D}_B$ does not remove this dependence. We compare $\boldsymbol{\tau}_A^\star$ to the in-sample $\boldsymbol{\tau}^\star$ on the synthetic protocol of Table 1 (d=20, $T\in\{100, 500\}$, acc $\in \{0.4, 0.6, 0.9\}$, 5 seeds).

Table 26: Cross-fitted vs. in-sample EB temperature. Mean absolute deviation $\mathrm{MAD}(\boldsymbol{\tau}^\star, \boldsymbol{\tau}_A^\star)$ averaged over edges and seeds; downstream AUROC of the resulting estimator.

T	acc	MAD($\boldsymbol{\tau}^\star, \boldsymbol{\tau}_A^\star$)	AUROC (in-sample)	AUROC (cross-fit)	Δ AUROC
100	0.4	0.038	.657 ± .015	.651 ± .018	−0.006
100	0.6	0.031	.698 ± .013	.694 ± .014	−0.004
100	0.9	0.022	.812 ± .011	.808 ± .013	−0.004
500	0.4	0.040	.813 ± .032	.811 ± .031	−0.002
500	0.6	0.029	.870 ± .013	.869 ± .014	−0.001
500	0.9	0.018	.939 ± .018	.940 ± .017	+0.001

MAD is at most 0.04 and downstream AUROC differs by at most 0.006 in this sweep. If confirmed from the matching records, these summaries would suggest limited sensitivity to this split for the tested configurations. They do not establish that in-sample agreement is unbiased or that a coherent wrong prior cannot reinforce itself. A stronger held-out analysis would need independently evaluated predictive criteria and appropriate temporal separation.

W Bilevel Fixed-Point Analysis

Let T_W and T_τ be exact local block-solution maps in neighborhoods where they are single-valued. If their Lipschitz constants are L_W and L_τ, their composition satisfies $\mathrm{Lip}(T_\tau \circ T_W) \le L_\tau L_W$. A contraction follows only when $q = L_\tau L_W < 1$ and the neighborhood is invariant. Under these assumptions the fixed point is locally unique.

With per-iteration map errors e_j, repeated application of the triangle inequality gives

$$\|z_t - z^\diamond\| \le q^t \|z_0 - z^\diamond\| + \sum_{j=0}^{t-1} q^{t-1-j} e_j.$$

For uniformly bounded error e, the asymptotic error bound is $e/(1-q)$. Smooth gradients alone do not establish positive curvature of the trust objective, and a positive ridge term is $\lambda_2\Omega$, not an unweighted $\lambda_2 I$. No numerical contraction factor is certified for the implemented alternating Adam routine.

Thresholded graph stability. Define $S_\beta(W) = \{r : |W_r| > \beta\|W_0\|_\infty\}$, matching the common instantaneous-scale threshold applied across lags. Let $m_\beta(W) = \min_r \big||W_r| - \beta\|W_0\|_\infty\big|$. Under a perturbation ΔW, every threshold score changes by at most $(1+\beta)\|\Delta W\|_\infty$. Thus the reported graph stays unchanged if $\|\Delta W\|_\infty < m_\beta(W)/(1+\beta)$. This sufficient margin certificate concerns a thresholded graph, not the exact ℓ_1 support or the active-coordinate hypotheses in Proposition 1.

Extended traces in Appendix AA.4 use the final five iterations to avoid counting a last-iteration observation as evidence of stability. Small metric variation and residuals are reported separately from this binary graph criterion.

X Bayesian Baseline (BayesDAG): Empirical Comparison

BayesDAG (Annadani et al., 2023) is a Bayesian cross-sectional baseline. The recorded suite uses its linear variant with 4 SG-MCMC chains, 100 training epochs, and 100-sample posterior averaging on CausalTime, cross-sectional synthetic data, and time-series synthetic data. It receives no external edge-probability matrix. A separate two-chain, 30-epoch static experiment is reported in Appendix AA.4; the two configurations are reported separately.

Setup. For CausalTime, PRCD-MAP uses the five cached data-driven priors, while BayesDAG uses its native no-external-prior model. Cross-sectional PRCD-MAP uses controlled-accuracy priors; the repeated BayesDAG values across accuracy levels are prior-independent. On temporal data, BayesDAG fits an instantaneous graph and PRCD-MAP can also model lags. Five seeds are used per setting except the recorded Medical BayesDAG baseline, which has four available runs, with $d = 20$ except AQI ($d = 36$).

Table 27: BayesDAG vs. PRCD-MAP across benchmark families. AUROC (mean±std). **Bold**: better of the two on each row. "W_0 only": instantaneous edges only (BayesDAG's native output); "combined": PRCD-MAP's full W_0+W_k output (BayesDAG cannot produce this).

Setting	BayesDAG (W_0)	PRCD-MAP (W_0)	PRCD-MAP (combined)
CausalTime real-world (T=400)			
AQI (d=36)	$.579 \pm .019$	$\mathbf{.612 \pm .069}$	$.693 \pm .060$
Medical (d=20)	$.507 \pm .011$	$\mathbf{.681 \pm .036}$	$.583 \pm .042$
Traffic (d=20)	$.540 \pm .017$	$\mathbf{.641 \pm .055}$	$.613 \pm .021$
Cross-sectional synthetic (d=20, acc=0.6)			
n=100	$.685 \pm .058$	$\mathbf{.713 \pm .123}$	—
n=500	$.756 \pm .058$	$\mathbf{.864 \pm .066}$	—
Cross-sectional synthetic (d=20, acc=0.9)			
n=100	$.685 \pm .058$	$\mathbf{.920 \pm .075}$	—
n=500	$.756 \pm .058$	$\mathbf{.947 \pm .050}$	—
Time-series synthetic (T=1000)			
Lorenz-96 (d=10)	$.577 \pm .009$	$.533 \pm .051$	$\mathbf{.905 \pm .024}$
Nonlinear SVAR (d=20)	$.784 \pm .073$	$.711 \pm .219$	$\mathbf{.821 \pm .180}$

Reading. The instantaneous-output comparison favors PRCD-MAP on the three CausalTime datasets and the listed static cells, with different prior information. BayesDAG is stronger on instantaneous Lorenz-96 (0.577 versus 0.533) and nonlinear SVAR (0.784 versus 0.711). PRCD-MAP's combined output has higher AUROC in those temporal examples, but it represents a different target. Recorded wall times are 80–210 seconds per BayesDAG seed and approximately 7 seconds for PRCD-MAP; differing model and optimization budgets prevent interpreting their ratio as a controlled computational advantage.

Y CausalTime Cached-Prior Pipeline: Statistical Significance

We assess whether the $+0.055$ "trust (cached)" vs. PCMCI+ improvement and the $+0.053$ "trust (cached)" vs. "no-prior" improvement (Table 2) are statistically significant. The trust (cached) row aggregates 5 priors per dataset (five cached data-driven constructions) $\times$ 5 seeds = 25 runs per dataset. Paired t-tests are computed across the 15 (prior, dataset) pairs (per-prior dataset means).

Of the five selected priors, four exceed the no-prior baseline on AQI and all five do so on Medical; three do so on Traffic. This describes the tested prior ensemble, not the distribution of all possible elicited priors. The reported aggregate $p = 0.008$ cannot be verified without the Traffic records. Priors from the same dataset share data and comparators, so the pooled analysis should be interpreted descriptively, not as evidence from 15 independent datasets. These unadjusted tests do not establish equivalence when nonsignificant or a universal robustness guarantee.

Relation to expected-risk safety. Definition 1 concerns expected predictive risk under a specified joint distribution. These finite-seed AUROC comparisons do not certify that property. A nonsignificant Traffic difference is not an equivalence test; the positive and negative prior-specific outcomes remain relevant limitations.

Table 28: Paired tests on CausalTime. Traffic and Aggregate rows are provisional and are not confirmed significance results. Δ = AUROC(trust, cached)−AUROC(comparator); p-values are paired t-test, two-sided. Per-dataset rows use five prior means; aggregate uses $5\times3 = 15$ prior–dataset means. External baseline means are treated as fixed, omitting their estimation uncertainty. These comparisons are conditional summaries over selected constructions.

Dataset	Comparison	Δ	p-value	Verdict
AQI	trust(cached) vs. PCMCI+	$+0.123$	0.014	sig. trust better
	trust(cached) vs. no-prior	$+0.067$	0.087	n.s.
	trust(cached) vs. BayesDAG	$+0.114$	0.018	sig. trust better
Medical	trust(cached) vs. PCMCI+	$+0.043$	0.075	n.s.
	trust(cached) vs. no-prior	$+0.089$	0.008	sig. trust better
	trust(cached) vs. BayesDAG	$+0.076$	0.013	sig. trust better
Traffic	trust(cached) vs. PCMCI+	-0.002	0.84	unverified
	trust(cached) vs. no-prior	$+0.002$	0.84	unverified
	trust(cached) vs. BayesDAG	$+0.073$	< 0.001	unverified
Aggregate	trust(cached) vs. PCMCI+	$+0.055$	0.008	unverified
	trust(cached) vs. no-prior	$+0.053$	0.012	unverified
	trust(cached) vs. BayesDAG	$+0.088$	< 0.001	unverified

Z Mechanism Decomposition, Realised Constants, and Practitioner Diagnostics

This section gives the full per-cell estimation protocol for the mechanism decomposition (Table 3 in the main text), realised constants for Theorem 5, the lagged-prior extension, the $\widehat{\rho}_{\text{cons}}$ practitioner test, and Traffic trust-attenuation diagnostics.

Z.1 CausalTime Mechanism Decomposition: Per-Cell Estimation Protocol

Evidence status. The component table is an exploratory, provisional allocation rather than a reproducible factorial decomposition. Its projection rule is unspecified and the complete direct-ablation records have not been located.

Table 3 (main text) decomposes the trust(cached)−PCMCI+ gap on CausalTime into four descriptive component estimates: (M1) soft-prior framework, (M2) EB calibration, (M3) per-edge MLP trust, (M4) cached-prior content. The component estimates combine the following protocols:

- **M1** from the trust(cached) row of Table 2 minus the no-prior row, projected onto the population mean of the EB-fixed-point $\boldsymbol{\tau}$.
- **M2** from the synthetic learned-vs-fixed gap of Table 7 restricted to the cached-prior accuracy regime ($\text{acc}\in[0.4, 0.6]$; synthetic gap $+0.020\pm0.018$). *Validity of transfer.* The synthetic gap is a structural-distribution-mismatched proxy; we cross-validated it with a direct CausalTime ablation that runs PRCD-MAP(trust, learned $\boldsymbol{\tau}$) versus PRCD-MAP(trust, fixed $\boldsymbol{\tau}=\mathbf{1}$) on the same 5 cached priors $\times$ 5 seeds (25 paired runs per dataset). The direct measurement gives a per-dataset M2 of $+0.030\pm0.040$ (AQI), $+0.047\pm0.021$ (Medical), -0.0004 ± 0.013 (Traffic), and an aggregate $+0.026\pm0.033$ across $n=75$ paired runs, sitting $+0.006$ AUROC *above* the transferred $+0.020$ point estimate. In this provisional summary, the learned-versus-fixed difference is largest on Medical (Medical $\gg$ AQI $\gg$ Traffic), but the Traffic estimate is unverified; anonymous variable names do not make a correlation prior uninformative. The transferred and direct estimates differ in design and cannot be treated as a certified lower bound or an independent additive effect.
- **M3** from the controlled-accuracy 10-seed trust-vs-per-group test of Table 6 (per-dataset Δ).
- **M4** from the prior-neutralization control of App. R that replaces the supplied prior with a uniform 0.5 baseline.

The displayed components approximately reproduce the aggregate gap, but use different priors, distributions, and reference conditions. They are not an orthogonal factorial decomposition; numerical agreement of their sum does not establish causal attribution or conservative error bars.

Z.2 Empirical Weak-Data Sweep ($T \ll d$ Regime)

We empirically characterize the small-sample behavior of the trust surrogate in Eq. (9). Setup: d=20, K=1, ER graph (edge prob. 0.15), Gaussian noise, controlled-flip prior, 5 seeds. We sweep $T \in \{20, 50, 100, 200, 500\}$ (so $T/d \in \{1, 2.5, 5, 10, 25\}$) and $\mathrm{acc} \in \{0.4, 0.6, 0.9\}$, comparing three variants: learned-$\boldsymbol{\tau}$ via the EB pipeline, fixed-$\boldsymbol{\tau}$=$\mathbf{1}$ (uncalibrated trust at $\widehat{P}$=P_{prior}), and low-trust ($\boldsymbol{\tau}$=$\boldsymbol{\tau}_{\min}$), which is only approximately neutral for a nonconstant prior.

Table 29: Weak-data sweep: AUROC (mean over 5 seeds) at d=20, K=1, ER graph, Gaussian noise. The T=20 column corresponds to T=d (the lowest sample ratio tested); learned-$\boldsymbol{\tau}$ collapses there to chance level (0.498) at acc=0.4 and underperforms the fixed-$\boldsymbol{\tau}$=$\mathbf{1}$ baseline at acc=0.9 (0.719 vs. 0.762). At accuracy 0.9, fixed trust remains stronger at $T = 100$ and $T = 200$; learned trust is stronger at $T = 500$. This sweep does not establish a universal sample-size boundary.

acc	Variant	T=20	T=50	T=100	T=200	T=500
0.4	learned-$\boldsymbol{\tau}$	.498	.588	.729	.820	.893
	fixed-$\boldsymbol{\tau}$=$\mathbf{1}$	.460	.546	.637	.712	.780
	low trust ($\boldsymbol{\tau}_{\min}$)	.540	.652	.768	.873	.932
0.6	learned-$\boldsymbol{\tau}$	.580	.687	.778	.860	.890
	fixed-$\boldsymbol{\tau}$=$\mathbf{1}$	.584	.676	.755	.829	.885
	low trust ($\boldsymbol{\tau}_{\min}$)	.557	.665	.775	.876	.933
0.9	learned-$\boldsymbol{\tau}$	.719	.792	.840	.910	.953
	fixed-$\boldsymbol{\tau}$=$\mathbf{1}$	.762	.814	.853	.916	.945
	low trust ($\boldsymbol{\tau}_{\min}$)	.597	.688	.784	.884	.934

Diagnostic on $\boldsymbol{\tau}^\star$. At $T = 20$, mean learned trust is approximately 0.53 for accuracy 0.4 and 1.61 for accuracy 0.9. The direction agrees with the accuracy ordering, but the graph metrics show that larger or smaller trust alone does not ensure better recovery. Fixed and low-trust variants remain informative comparison points; these data do not isolate a unique cause of failure.

Z.3 Realised Penalty Ranges and Conditioning

For $d = 20, K = 1, T = 500$, the recorded instantaneous penalty distribution at accuracy 0.6, averaged over 10 seeds, has mean 0.94, median 0.95, minimum 0.51, and maximum 1.50. The observed ratio is $1.50/0.51 \approx 2.94$. The exact sigmoid range gives a structural floor of 0.5; clipping the raw prior at $1 - 10^{-3}$ does not impose the same upper bound after temperature sharpening.

These observations describe penalty variation, not a measured restricted-eigenvalue constant or a certified risk bound. In particular, a factor 3 in penalty weights does not by itself give a factor 3, 6, or 4 in estimation risk. A covariance eigenvalue of 0.4, where observed, must still be distinguished from the residual-weighted Huber curvature with $1/(Td)$ normalization. The empirical learned-versus-fixed AUROC gap is a performance measurement and cannot be inserted as a theoretical constant to certify that same gap.

Z.4 Lagged-Prior Semantics

The framework of §3 applies the same per-entry prior $P_{\mathrm{prior},ij}$ to both $W_{0,ij}$ (instantaneous) and $W_{k,ij}, k{\geq}1$ (lagged) entries via the same mask $\boldsymbol{\Omega}(\boldsymbol{\tau})$ in Eq. (2). The prior refers to a dependency at any lag, as specified below.

What $\mathbf{P}_{\mathrm{prior}}$ encodes. Expert and LLM elicitation of causal priors typically yields lag-agnostic edge confidence: "does i causally influence j?", not "does i influence j with lag k=3?". Our parameterization respects this: $P_{\mathrm{prior},ij}$ is the elicited confidence in the existence of *any* directed

dependency $i \to j$, whether contemporaneous or lagged. Applying the same prior to all $W_{k,ij}$ then encodes the assumption: *the same edge-confidence value regularizes every lag coefficient*. This repeats the regularization information across lags; it is not a normalized probability allocation over mutually exclusive lag locations.

Modeling choice. When supplied knowledge concerns a union of instantaneous and lagged dependencies, reusing its score across lags is one regularization choice. Such a union score does not establish that every lag coefficient should receive the same prior strength.

When lag-resolved priors are available. One can replace the shared mask by $\Omega^{(k)}$ and use lag-specific temperatures. This changes the prior model and the trust-update design; the lag coefficients are jointly fitted through a shared residual rather than independent Lasso problems. The local perturbation calculation extends after redefining the masks, but statistical and calibration guarantees require their own assumptions.

Empirical sensitivity to the assumption. The archived five-seed experiment uses $d = 20, T = 500, K = 1$, instantaneous prior accuracy 0.7, and lagged prior accuracy 0.4. Mean AUROC is 0.918 with the instantaneous prior shared across lags, 0.903 with the lagged prior shared across lags, and 0.926 with lag-resolved masks. The nominal no-prior ablation reaches 0.930. The lag-resolved gain over the shared instantaneous prior is 0.008; it is an empirical comparison for these settings, not a general guarantee.

Z.5 Practitioner Heterogeneity Test for Trust Propagation

Evidence status. The numerical proxy values below have not been matched to reproducible per-scenario output records.

The usefulness of contextual trust depends on the data and prior. The following inexpensive covariance-based diagnostic is an exploratory heuristic for choosing architectures. It is not a certified predictor of downstream gain, and the matched comparison in Appendix AA.3 favors grouped trust on aggregate.

Test statistic. Let $\mathbf{X} \in \mathbb{R}^{T \times d}$ denote the data and $\mathbf{P}_{\text{prior}} \in [0,1]^{d \times d}$ the prior. Compute the *neighborhood-consistency proxy*

$$\widehat{\rho}_{\text{cons}}(\mathbf{P}_{\text{prior}}, \mathbf{X}) := \text{med}_{(i,j)} \, \text{Corr}\Big(\mathbf{P}_{\text{prior},\mathcal{N}(i,j)}, \; |\widehat{\mathbf{\Sigma}}|_{\mathcal{N}(i,j)}\Big),$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance of $\mathbf{X}$ (cheap, $O(Td^2)$). $\widehat{\rho}_{\text{cons}}$ uses sample covariance rather than true edge labels, so it is not a consistent estimator of prior-versus-ground-truth neighborhood correlation without additional assumptions; the median also differs from a minimum over neighborhoods. It is computable without running PRCD-MAP.

Exploratory decision rule. A threshold of 0.20 was examined on the scenarios below. Traffic has proxy 0.22 and gain 0.013, so these observations do not support a guarantee of gain at least 0.015 above the threshold. Out-of-sample validation is required before using this heuristic as an architecture-selection rule.

Table 30: Practitioner heterogeneity test: $\widehat{\rho}_{\text{cons}}$ vs. realised trust-propagation gain.

Scenario	$\widehat{\rho}_{\text{cons}}$	Gain (trust−per-group)	Decision
Synthetic ER, iid corruption (App. O)	0.04	+0.001	skip MLP
BA hub–peripheral, acc(.95, .20)	0.41	+0.030	enable MLP
CausalTime AQI (mixed protocols)	0.27	+0.031	enable MLP
CausalTime Medical (mixed protocols)	0.31	+0.039	enable MLP
CausalTime Traffic (mixed protocols)	0.22	+0.013	enable (marginal)

Interpretation. The proxy is cheap to compute and broadly tracks the listed examples, but correlations can reflect common causes, indirect paths, or confounding. It does not identify causal prior correctness from observational inputs. The CausalTime gains come from the controlled-accuracy sweep, whereas the listed proxies concern cached priors; these unmatched conditions cannot validate the decision rule.

Z.6 Traffic Cached-Prior Trust Attenuation: Per-Prior $\tau^\star$ Distribution

Evidence status. The Traffic trust quantiles and prior-accuracy values below lack matched saved temperature arrays or per-run records. They are provisional and cannot support a confirmed calibration claim.

Section 4.3.1 reports a Traffic difference of $+0.002$ between trust(cached) and no-prior, with a provisional $p = 0.84$. The per-prior trust distribution below has not been verified from saved arrays. Without the underlying arrays these quantities cannot establish attenuation, equivalence, or a calibration theorem.

Table 31: Provisional mean and inter-quantile range (saved arrays unavailable) of the EB-learned $\tau^\star$ across the 5 cached priors on Traffic (T=400, 5 seeds per prior). $\tau^\star$ is averaged over off-diagonal edges; $\tau_{\min}$=10^{-3}, $\tau_{\max}$=2.

Prior identifier	$\widehat{\tau^\star}$ (mean)	[q25, q75]	Empirical prior accuracy
Pearson	0.043	[0.012, 0.072]	0.51
Spearman	0.067	[0.018, 0.103]	0.53
Sparse Pearson	0.038	[0.011, 0.062]	0.49
Squared Pearson	0.054	[0.014, 0.085]	0.50
Partial correlation	0.029	[0.008, 0.048]	0.48
Mean across 5 priors	0.046	—	0.50

Reading. Mean trust across the five priors ranges from 0.029 to 0.067, with overall mean 0.046. This is about 37 times smaller than the synthetic high-accuracy mean of 1.7, yet remains above $\tau_{\min} = 0.001$. This is only a description of the provisional table, not a verified attenuation result.

AA Controlled Priors and Extended Diagnostic Experiments

We evaluate three questions: whether informative priors help recovery, whether learned trust suppresses structured errors, and whether a contextual module is necessary. Tables are generated from per-run files. Detailed configurations, inputs, scripts, and outputs accompany the manuscript.

Metrics and information regime. AUROC evaluates off-diagonal absolute edge scores. The evaluator chooses the best F1 over 200 thresholds using the evaluation graph; we denote this $\mathrm{F1}_{\mathrm{oracle}}$. Reported $\mathrm{SHD}_{\mathrm{oracle}}$ is directed adjacency Hamming distance at that same threshold, counting a reversed edge as two entry errors. These are retrospective diagnostics, not label-free deployment performance. Full per-condition summaries also include AUPRC and standard deviations. Priors in the controlled experiments are generated from the reference graph at specified corruption levels, including on real measured data. They test prior consumption, not independently elicited expert accuracy. Nominal and realized accuracies differ for structured corruption and are both recorded.

AA.1 CausalTime and cross-sectional recovery

CausalTime provides realistically generated time series based on real datasets (Cheng et al., 2024); its reference graph is a benchmark target, not a newly established causal ground truth for the original application. We evaluate the collapsed instantaneous-plus-lagged score. The collapsed target can be cyclic even when the within-slice graph is acyclic; we do not report standard DAG SID for this target. Table 32 reports five-seed results at controlled prior accuracy 0.9. Informative priors improve ranking and structural diagnostics, but baselines do not receive the extra reference-derived information. At accuracy 0.6, Medical learned-trust AUROC is 0.535 versus PCMCI+ at 0.540,

Table 32: CausalTime: five-seed means at controlled accuracy 0.9 for prior-consuming methods. PCMCI+ receives no external prior. Structural metrics use oracle thresholds.

Dataset	Method	AUROC	$F1_{oracle}$	SHD_{oracle}
AQI	PCMCI+	0.570	0.403	942.0
AQI	PRCD contextual	0.815	0.687	190.2
AQI	PRCD grouped	0.702	0.498	382.8
Traffic	PCMCI+	0.615	0.336	166.0
Traffic	PRCD contextual	0.688	0.396	87.6
Traffic	PRCD grouped	0.682	0.381	115.4
Medical	PCMCI+	0.540	0.530	227.0
Medical	PRCD contextual	0.661	0.556	178.8
Medical	PRCD grouped	0.585	0.521	238.2

Table 33: Structured-error suite: mean over 80 matched conditions per method. All structural metrics use oracle thresholds.

Method	AUROC	$F1_{oracle}$	SHD_{oracle}
Learned edge trust	0.850	0.694	36.0
Fixed trust	0.863	0.714	33.2
No prior	0.759	0.544	62.4

while AQI learned and grouped variants both obtain 0.597. The gains are therefore neither monotone for every finite-sample condition nor uniform across datasets. All prior levels and baseline rows are included in Appendix AB.

On Sachs (Sachs et al., 2005) and CausalChambers (Gamella et al., 2025) measured data, the same controlled-prior protocol tests cross-sectional estimation ($K = 0$). At accuracy 0.9, learned trust obtains AUROC/$F1_{oracle}$/SHD_{oracle} of 0.689/0.409/32.4 and 0.724/0.352/62.8, respectively. These improvements are conditional on supplying an informative reference-derived prior. At accuracy 0.4 on Sachs, learned trust gives AUROC 0.570 but no F1 improvement over NOTEARS. The full table also reports the lower-accuracy settings.

AA.2 Structured prior errors

The structured-error suite covers block misconnections, direction flips, group omissions, and LLM-style false-positive corruption, with $T \in \{100, 500\}$, nominal accuracies 0.4 and 0.6, and five seeds. The three methods share 80 conditions, giving 240 fits. "LLM-style" here names a synthetic error family, not an actual language-model run. The no-prior ablation sets $P = 1/2$, disables trust learning, and sets $\lambda_2 = 0$. It therefore differs from the same-regularization neutral estimator in Proposition 1; its comparison also includes the effect of ridge regularization and cannot isolate calibration alone. Learned and fixed trust provide the closer comparison of adaptive versus fixed consumption.

Learned trust ranks correct versus wrong assertions with mean AUROC 0.796. Mean temperatures are 1.641 and 0.661, respectively, and wrong-assertion confidence moves toward 0.5 by 0.232 on average; attenuation is 0.385 in the false-positive hallucination family. Correctness labels are used only for evaluation. These measurements directly support selective attenuation. Nevertheless, fixed trust has higher mean downstream AUROC and F1 (Table 33). Ranking prior reliability and maximizing graph recovery are different objectives, and an internally coherent wrong prior may evade detection when data are weak.

AA.3 Matched architecture comparison

The matched suite compares edge-GAT, lightweight MLP, grouped temperatures, and fixed trust across the same 40 conditions per method. Grouped calibration has the highest aggregate AUROC and F1 (Table 34). Thus the six-feature MLP is an optional contextual extension. Its extra flexibility may help individual conditions but is not required for useful prior consumption. Aggregate standard

Table 34: Matched architecture suite: mean over 40 conditions per method. Grouped trust is strongest on average in ranking and oracle F1.

Method	AUROC	$F1_{oracle}$	SHD_{oracle}
Edge-GAT	0.903	0.788	6.5
Lightweight MLP	0.898	0.783	6.4
Grouped trust	0.907	0.804	5.8
Fixed trust	0.905	0.794	6.5

deviations across heterogeneous conditions are descriptive, not confidence intervals for a common population mean.

AA.4 Optimizer stability and Bayesian comparison

Across 60 extended optimizer runs, 42 have an unchanged thresholded graph over the final five iterations. The graph is defined by a relative magnitude threshold, not exact nonzero support. Across all runs, mean final-ten-iteration AUROC and F1 ranges are 0.000889 and 0.003363; for the 18 unstable runs they are 0.00185 and 0.00595. The mean final DAG residual is about 10^{-6}. These observations suggest limited metric sensitivity to late threshold crossings, but do not verify exact active-set stability or global convergence.

The BayesDAG comparison uses a static acyclic $d = 20$ SEM, two chains, 30 epochs, and five seeds at each sample size. At $n = 100, 500$, BayesDAG AUROC is 0.597, 0.653; PRCD with fixed trust and accuracy 0.6 gives 0.761, 0.842. BayesDAG receives no external prior. This comparison assesses the methods in different information regimes; it does not isolate how they consume identical priors.

AB Complete Controlled-Experiment Summaries

The following tables are computed directly from per-run results. Means and sample standard deviations are calculated within each dataset, method, and prior level. A dash denotes a method without the controlled external prior. Standard deviations of zero can arise for deterministic algorithms applied to fixed input data. The high-level tables in the main text are selected views of these same files. $F1_{oracle}$ and SHD_{oracle} retain the retrospective threshold convention described in Section AA.

Table 35: CausalTime AQI: mean and sample standard deviation over five seeds.

Dataset	Method	Prior	AUROC	AUPRC	$F1_{oracle}$	SHD_{oracle}
AQI	DYNOTEARS	–	0.554±0.011	0.375±0.004	0.403±0.000	942.0±0.0
AQI	PCMCI+	–	0.570±0.000	0.396±0.000	0.403±0.000	942.0±0.0
AQI	VARLiNGAM	–	0.541±0.000	0.306±0.000	0.403±0.000	942.0±0.0
AQI	PRCD contextual	0.3	0.620±0.001	0.353±0.001	0.430±0.001	573.8±17.6
AQI	PRCD grouped	0.3	0.624±0.001	0.358±0.001	0.434±0.001	583.8±8.4
AQI	PRCD contextual	0.6	0.597±0.010	0.368±0.007	0.410±0.007	742.2±161.6
AQI	PRCD grouped	0.6	0.597±0.010	0.368±0.007	0.410±0.007	742.2±161.6
AQI	PRCD contextual	0.9	0.815±0.015	0.694±0.029	0.687±0.014	190.2±17.0
AQI	PRCD grouped	0.9	0.702±0.008	0.526±0.010	0.498±0.020	382.8±39.3

Table 36: CausalTime Traffic: mean and sample standard deviation over five seeds.

Dataset	Method	Prior	AUROC	AUPRC	$F1_{oracle}$	SHD_{oracle}
Traffic	DYNOTEARS	–	0.583 ± 0.013	0.207 ± 0.003	0.316 ± 0.013	149.2 ± 41.8
Traffic	PCMCI+	–	0.615 ± 0.000	0.258 ± 0.000	0.336 ± 0.000	166.0 ± 0.0
Traffic	VARLiNGAM	–	0.458 ± 0.000	0.170 ± 0.000	0.281 ± 0.000	318.0 ± 0.0
Traffic	PRCD contextual	0.3	0.599 ± 0.004	0.269 ± 0.011	0.335 ± 0.003	147.8 ± 5.1
Traffic	PRCD grouped	0.3	0.574 ± 0.015	0.241 ± 0.016	0.317 ± 0.016	167.2 ± 14.4
Traffic	PRCD contextual	0.6	0.616 ± 0.015	0.297 ± 0.016	0.343 ± 0.012	148.6 ± 8.4
Traffic	PRCD grouped	0.6	0.612 ± 0.012	0.291 ± 0.023	0.337 ± 0.011	155.6 ± 11.5
Traffic	PRCD contextual	0.9	0.688 ± 0.019	0.396 ± 0.023	0.396 ± 0.020	87.6 ± 27.2
Traffic	PRCD grouped	0.9	0.682 ± 0.008	0.400 ± 0.021	0.381 ± 0.010	115.4 ± 27.8

Table 37: CausalTime Medical: mean and sample standard deviation over five seeds.

Dataset	Method	Prior	AUROC	AUPRC	$F1_{oracle}$	SHD_{oracle}
Medical	DYNOTEARS	–	0.523 ± 0.006	0.415 ± 0.003	0.519 ± 0.000	247.0 ± 0.0
Medical	PCMCI+	–	0.540 ± 0.000	0.412 ± 0.000	0.530 ± 0.000	227.0 ± 0.0
Medical	VARLiNGAM	–	0.505 ± 0.000	0.370 ± 0.000	0.519 ± 0.000	247.0 ± 0.0
Medical	PRCD contextual	0.3	0.511 ± 0.003	0.353 ± 0.001	0.520 ± 0.001	245.6 ± 1.3
Medical	PRCD grouped	0.3	0.497 ± 0.008	0.343 ± 0.005	0.519 ± 0.000	245.8 ± 2.7
Medical	PRCD contextual	0.6	0.535 ± 0.008	0.382 ± 0.005	0.521 ± 0.003	243.4 ± 5.7
Medical	PRCD grouped	0.6	0.538 ± 0.007	0.382 ± 0.006	0.519 ± 0.001	237.2 ± 21.9
Medical	PRCD contextual	0.9	0.661 ± 0.055	0.549 ± 0.083	0.556 ± 0.043	178.8 ± 58.5
Medical	PRCD grouped	0.9	0.585 ± 0.007	0.428 ± 0.006	0.521 ± 0.004	238.2 ± 14.3

Table 38: Sachs: five-seed mean and sample standard deviation. Priors are reference-derived.

Dataset	Method	Prior	AUROC	AUPRC	$F1_{oracle}$	SHD_{oracle}
sachs	NOTEARS	–	0.512 ± 0.026	0.215 ± 0.007	0.308 ± 0.000	90.0 ± 0.0
sachs	DAGMA	–	0.517 ± 0.011	0.220 ± 0.002	0.308 ± 0.000	90.0 ± 0.0
sachs	PRCD contextual	0.4	0.570 ± 0.070	0.227 ± 0.035	0.308 ± 0.000	90.0 ± 0.0
sachs	PRCD grouped	0.4	0.573 ± 0.033	0.221 ± 0.028	0.308 ± 0.000	90.0 ± 0.0
sachs	PRCD contextual	0.6	0.602 ± 0.039	0.255 ± 0.052	0.323 ± 0.021	57.6 ± 29.6
sachs	PRCD grouped	0.6	0.602 ± 0.065	0.242 ± 0.040	0.313 ± 0.011	79.2 ± 24.1
sachs	PRCD contextual	0.9	0.689 ± 0.027	0.355 ± 0.033	0.409 ± 0.038	32.4 ± 1.9
sachs	PRCD grouped	0.9	0.673 ± 0.015	0.302 ± 0.005	0.333 ± 0.006	36.0 ± 1.0

Table 39: Causalchambers: five-seed mean and sample standard deviation. Priors are reference-derived.

Dataset	Method	Prior	AUROC	AUPRC	$F1_{oracle}$	SHD_{oracle}
causalchambers	NOTEARS	–	0.405 ± 0.022	0.108 ± 0.014	0.192 ± 0.009	174.0 ± 152.5
causalchambers	DAGMA	–	0.494 ± 0.024	0.128 ± 0.010	0.204 ± 0.012	119.8 ± 123.7
causalchambers	PRCD contextual	0.4	0.666 ± 0.032	0.186 ± 0.031	0.241 ± 0.026	74.4 ± 8.2
causalchambers	PRCD grouped	0.4	0.640 ± 0.045	0.174 ± 0.040	0.224 ± 0.035	131.4 ± 117.3
causalchambers	PRCD contextual	0.6	0.681 ± 0.030	0.202 ± 0.017	0.258 ± 0.049	68.8 ± 6.7
causalchambers	PRCD grouped	0.6	0.667 ± 0.020	0.198 ± 0.021	0.245 ± 0.040	77.6 ± 5.6
causalchambers	PRCD contextual	0.9	0.724 ± 0.022	0.284 ± 0.022	0.352 ± 0.015	62.8 ± 7.3
causalchambers	PRCD grouped	0.9	0.714 ± 0.023	0.270 ± 0.025	0.325 ± 0.047	68.0 ± 4.8

AC Reproducibility Details

The supplementary package contains model code, experiment entry points, input data, numerical outputs, dependency specifications, and third-party source with its license notices. The table generator

validates numeric inputs and produces both CSV summaries and LaTeX tables. Per-run results retain their numerical precision.

Table 40: Experiment entry points in the supplementary package. Each script is under `code/`; numerical outputs are under the corresponding `results/` directory.

Experiment	Entry point
Real cross-sectional data	`real_world.py`
CausalTime	`causaltime.py`
Matched trust architectures	`architecture_comparison.py`
Optimizer stability	`stability.py`
Structured prior errors	`structured_errors.py`
Static DAG comparison	`static_dag.py`

From the supplementary root, `python scripts/build_tables.py` generates the tables without fitting models. On Linux with CUDA, `bash scripts/bootstrap.sh` creates the numerical environments. Experiments can be run with `bash scripts/run_experiment.sh structured_errors 0`, where the final argument selects the GPU. Supported experiment names match the entry points above. The static DAG runner also invokes the BayesDAG baseline in a separate environment. Dependency versions and per-experiment configuration records accompany the code.

AD Complete Component, Quality, and Correlated-Error Sweeps

These experiments use the shared linear SVAR generator with $d = 20$, $T = 500$, $K = 1$, instantaneous ER edge probability 0.15, lagged probability 0.10, and five seeds unless varied explicitly. The contextual runner uses $\lambda_1 = 0.001$, $\lambda_2 = 0.01$, trust bounds $[0.05, 3]$, 35 outer iterations, up to 400 inner steps, and eight trust steps. These overrides differ from some of the settings elsewhere in the paper and are retained in the configuration files. All tables below use per-condition means and sample standard deviations. AUROC and AUPRC concern the collapsed temporal graph; F1 and SHD use the evaluation-label-selected threshold. Reversed edges count as two adjacency errors. The per-run files additionally retain instantaneous-only and top-k diagnostics.

AD.1 Component objectives

The 80 fits compare the full contextual objective with no trust updates, no curvature term, fixed trust, grouped trust, and no prior. The no-update variant keeps its initialized contextual network; it is not identical to fixed unit trust. The no-curvature variant retains agreement, coefficient penalties, and network regularization. The no-prior variant uses a uniform probability matrix and also disables ridge, so its contrast is not a calibration-only effect. Its single row is prior-independent.

Table 41: Complete component-objective sweep.

Variant	Acc.	n	AUROC	AUPRC	F1	SHD
Fixed trust	0.3	5	0.716 ± 0.046	0.344 ± 0.049	0.471 ± 0.057	75.2 ± 12.9
Full	0.3	5	0.830 ± 0.050	0.567 ± 0.082	0.639 ± 0.050	45.4 ± 7.3
No trust updates	0.3	5	0.832 ± 0.047	0.574 ± 0.081	0.642 ± 0.054	43.0 ± 5.0
No curvature term	0.3	5	0.834 ± 0.048	0.581 ± 0.079	0.647 ± 0.053	43.8 ± 5.1
No prior	—	5	0.848 ± 0.056	0.644 ± 0.100	0.672 ± 0.062	40.2 ± 9.7
Grouped trust	0.3	5	0.857 ± 0.037	0.624 ± 0.051	0.678 ± 0.037	39.2 ± 5.3
Fixed trust	0.6	5	0.865 ± 0.035	0.636 ± 0.070	0.677 ± 0.065	37.4 ± 8.8
Full	0.6	5	0.874 ± 0.028	0.656 ± 0.063	0.707 ± 0.055	37.4 ± 8.3
No trust updates	0.6	5	0.866 ± 0.035	0.642 ± 0.076	0.682 ± 0.068	38.8 ± 10.9
No curvature term	0.6	5	0.874 ± 0.028	0.656 ± 0.063	0.709 ± 0.056	36.8 ± 8.2
Grouped trust	0.6	5	0.882 ± 0.019	0.704 ± 0.045	0.730 ± 0.044	31.8 ± 5.3
Fixed trust	0.9	5	0.949 ± 0.014	0.911 ± 0.027	0.886 ± 0.027	13.2 ± 3.8

Continued on next page

Continued

Variant	Acc.	n	AUROC	AUPRC	F1	SHD
Full	0.9	5	0.945 ± 0.018	0.903 ± 0.030	0.875 ± 0.033	14.4 ± 4.4
No trust updates	0.9	5	0.954 ± 0.010	0.917 ± 0.022	0.894 ± 0.022	12.2 ± 2.9
No curvature term	0.9	5	0.945 ± 0.018	0.903 ± 0.030	0.874 ± 0.033	14.4 ± 4.4
Grouped trust	0.9	5	0.944 ± 0.018	0.902 ± 0.030	0.879 ± 0.028	13.8 ± 3.6

AD.2 Prior-quality endpoints and noise

The 130 fits cover Gaussian and Laplace noise with prior accuracies $0, 0.2, 0.4, 0.6, 0.8, 1$. Each noise condition has one prior-independent no-prior comparator. Accuracy zero is a fully incorrect prior, not a zero-error endpoint; accuracy one does not remove finite-sample or optimization error. These curves measure robustness empirically and do not certify expected-risk safety.

Table 42: Complete prior-quality sweep; five seeds per row.

Noise	Method	Acc.	AUROC	F1	SHD
gaussian	No prior	—	0.848 ± 0.056	0.672 ± 0.062	40.2 ± 9.7
gaussian	Fixed trust	0.0	0.533 ± 0.032	0.299 ± 0.009	165.8 ± 72.7
gaussian	Full	0.0	0.812 ± 0.042	0.618 ± 0.030	47.4 ± 9.4
gaussian	Fixed trust	0.2	0.626 ± 0.029	0.361 ± 0.015	108.8 ± 29.8
gaussian	Full	0.2	0.792 ± 0.097	0.586 ± 0.116	59.4 ± 38.7
gaussian	Fixed trust	0.4	0.798 ± 0.035	0.579 ± 0.032	57.6 ± 9.5
gaussian	Full	0.4	0.833 ± 0.060	0.656 ± 0.080	43.0 ± 14.1
gaussian	Fixed trust	0.6	0.865 ± 0.035	0.677 ± 0.065	37.4 ± 8.8
gaussian	Full	0.6	0.874 ± 0.028	0.707 ± 0.055	37.4 ± 8.3
gaussian	Fixed trust	0.8	0.921 ± 0.034	0.823 ± 0.057	20.6 ± 6.7
gaussian	Full	0.8	0.912 ± 0.035	0.820 ± 0.065	20.8 ± 7.3
gaussian	Fixed trust	1.0	0.956 ± 0.017	0.897 ± 0.025	11.8 ± 3.2
gaussian	Full	1.0	0.954 ± 0.016	0.895 ± 0.022	12.0 ± 2.2
laplace	No prior	—	0.920 ± 0.022	0.790 ± 0.049	24.4 ± 6.7
laplace	Fixed trust	0.0	0.530 ± 0.053	0.302 ± 0.016	172.8 ± 85.1
laplace	Full	0.0	0.910 ± 0.040	0.786 ± 0.077	25.8 ± 10.7
laplace	Fixed trust	0.2	0.646 ± 0.062	0.398 ± 0.047	91.4 ± 12.9
laplace	Full	0.2	0.885 ± 0.106	0.764 ± 0.157	27.2 ± 19.4
laplace	Fixed trust	0.4	0.831 ± 0.049	0.647 ± 0.090	43.2 ± 14.6
laplace	Full	0.4	0.864 ± 0.089	0.716 ± 0.131	36.6 ± 21.0
laplace	Fixed trust	0.6	0.904 ± 0.019	0.767 ± 0.022	27.6 ± 3.4
laplace	Full	0.6	0.907 ± 0.021	0.787 ± 0.028	25.8 ± 4.1
laplace	Fixed trust	0.8	0.945 ± 0.019	0.857 ± 0.023	16.6 ± 2.7
laplace	Full	0.8	0.939 ± 0.024	0.854 ± 0.032	17.4 ± 4.2
laplace	Fixed trust	1.0	0.962 ± 0.014	0.911 ± 0.020	10.4 ± 2.7
laplace	Full	1.0	0.956 ± 0.017	0.906 ± 0.024	11.0 ± 3.2

AD.3 Trust as a detector of incorrect assertions

The 30 fits measure whether temperatures distinguish correct from incorrect prior assertions. The two accuracy endpoints contain only one correctness class, so trust AUROC is undefined there; the table uses a dash instead of the evaluator's numerical sentinel. Missing correct/wrong subgroup means also use dashes. Higher trust AUROC measures reliability ranking, not calibration of probabilities or guaranteed graph improvement.

Table 43: Trust discrimination and downstream recovery; five seeds per row.

Acc.	Trust AUROC	τ correct	τ wrong	Graph AUROC	Graph F1
0.0	—	—	0.054 ± 0.001	0.812 ± 0.042	0.618 ± 0.030
0.2	0.867 ± 0.029	0.273 ± 0.427	0.096 ± 0.092	0.792 ± 0.097	0.586 ± 0.116
0.4	0.869 ± 0.012	0.846 ± 0.417	0.182 ± 0.075	0.833 ± 0.060	0.656 ± 0.080

Continued on next page

Continued

Acc.	Trust AUROC	τ correct	τ wrong	Graph AUROC	Graph F1
0.6	0.873 ± 0.018	1.496 ± 0.335	0.299 ± 0.068	0.874 ± 0.028	0.707 ± 0.055
0.8	0.876 ± 0.028	1.736 ± 0.232	0.307 ± 0.086	0.912 ± 0.035	0.820 ± 0.065
1.0	—	2.067 ± 0.180	—	0.954 ± 0.016	0.895 ± 0.022

AD.4 Correlated and systematic errors

The 230 fits cover IID corruption, block misconnections, direction flips, group omissions, and synthetic false-positive errors, with nominal accuracies 0.4, 0.6, and 0.8. Full contextual, fixed, and grouped trust share each generated prior. The no-prior comparator is generated once per seed, yielding five additional fits. Realized accuracy is shown because structured construction can differ from its nominal target. This experiment and the 240-fit selective-attenuation suite use distinct grids and should not be pooled as duplicate observations.

Table 44: Complete correlated-error sweep; five seeds per row.

Error	Nom.	Method	Realized	AUROC	F1	SHD
Block errors	0.4	Fixed trust	0.842 ± 0.000	0.937 ± 0.025	0.845 ± 0.021	17.8 ± 2.9
Block errors	0.4	Full	0.842 ± 0.000	0.934 ± 0.026	0.849 ± 0.044	17.6 ± 6.1
Block errors	0.4	Grouped trust	0.842 ± 0.000	0.916 ± 0.048	0.812 ± 0.071	22.6 ± 11.7
Block errors	0.6	Fixed trust	0.895 ± 0.000	0.943 ± 0.019	0.859 ± 0.033	16.6 ± 5.0
Block errors	0.6	Full	0.895 ± 0.000	0.939 ± 0.019	0.856 ± 0.031	16.8 ± 3.9
Block errors	0.6	Grouped trust	0.895 ± 0.000	0.935 ± 0.021	0.844 ± 0.045	18.8 ± 6.9
Block errors	0.8	Fixed trust	0.947 ± 0.000	0.949 ± 0.017	0.873 ± 0.021	15.0 ± 3.3
Block errors	0.8	Full	0.947 ± 0.000	0.939 ± 0.017	0.859 ± 0.025	16.2 ± 3.6
Block errors	0.8	Grouped trust	0.947 ± 0.000	0.939 ± 0.017	0.860 ± 0.015	16.0 ± 2.3
Direction flips	0.4	Fixed trust	0.822 ± 0.015	0.759 ± 0.042	0.546 ± 0.052	58.0 ± 13.1
Direction flips	0.4	Full	0.822 ± 0.015	0.754 ± 0.029	0.541 ± 0.047	60.2 ± 15.6
Direction flips	0.4	Grouped trust	0.822 ± 0.015	0.769 ± 0.024	0.565 ± 0.037	54.0 ± 9.6
Direction flips	0.6	Fixed trust	0.879 ± 0.007	0.853 ± 0.035	0.660 ± 0.069	40.6 ± 10.8
Direction flips	0.6	Full	0.879 ± 0.007	0.835 ± 0.034	0.638 ± 0.073	46.6 ± 14.8
Direction flips	0.6	Grouped trust	0.879 ± 0.007	0.849 ± 0.037	0.660 ± 0.070	42.4 ± 16.3
Direction flips	0.8	Fixed trust	0.940 ± 0.006	0.887 ± 0.020	0.756 ± 0.038	29.4 ± 5.2
Direction flips	0.8	Full	0.940 ± 0.006	0.877 ± 0.024	0.748 ± 0.035	30.2 ± 4.6
Direction flips	0.8	Grouped trust	0.940 ± 0.006	0.878 ± 0.023	0.744 ± 0.039	31.0 ± 4.8
Group omissions	0.4	Fixed trust	0.955 ± 0.010	0.945 ± 0.020	0.864 ± 0.024	15.6 ± 3.3
Group omissions	0.4	Full	0.955 ± 0.010	0.944 ± 0.018	0.871 ± 0.019	14.6 ± 1.5
Group omissions	0.4	Grouped trust	0.955 ± 0.010	0.938 ± 0.015	0.855 ± 0.018	16.4 ± 2.8
Group omissions	0.6	Fixed trust	0.971 ± 0.007	0.946 ± 0.019	0.878 ± 0.029	14.6 ± 3.8
Group omissions	0.6	Full	0.971 ± 0.007	0.940 ± 0.018	0.865 ± 0.038	15.4 ± 4.2
Group omissions	0.6	Grouped trust	0.971 ± 0.007	0.939 ± 0.015	0.860 ± 0.038	16.0 ± 4.5
Group omissions	0.8	Fixed trust	0.985 ± 0.004	0.953 ± 0.017	0.889 ± 0.025	12.8 ± 2.7
Group omissions	0.8	Full	0.985 ± 0.004	0.946 ± 0.019	0.876 ± 0.037	14.0 ± 3.7
Group omissions	0.8	Grouped trust	0.985 ± 0.004	0.944 ± 0.022	0.869 ± 0.034	14.8 ± 3.7
IID	0.4	Fixed trust	0.400 ± 0.000	0.720 ± 0.031	0.487 ± 0.059	73.8 ± 26.3
IID	0.4	Full	0.400 ± 0.000	0.811 ± 0.094	0.625 ± 0.128	46.2 ± 22.2
IID	0.4	Grouped trust	0.400 ± 0.000	0.852 ± 0.069	0.693 ± 0.079	36.8 ± 9.3
IID	0.6	Fixed trust	0.600 ± 0.000	0.864 ± 0.054	0.679 ± 0.061	40.6 ± 9.3
IID	0.6	Full	0.600 ± 0.000	0.866 ± 0.052	0.687 ± 0.071	39.6 ± 10.5
IID	0.6	Grouped trust	0.600 ± 0.000	0.885 ± 0.039	0.744 ± 0.051	29.4 ± 6.1
IID	0.8	Fixed trust	0.800 ± 0.000	0.926 ± 0.047	0.803 ± 0.052	23.4 ± 6.3
IID	0.8	Full	0.800 ± 0.000	0.917 ± 0.047	0.788 ± 0.051	24.6 ± 6.1
IID	0.8	Grouped trust	0.800 ± 0.000	0.924 ± 0.024	0.809 ± 0.013	22.2 ± 2.8
False positives	0.4	Fixed trust	0.949 ± 0.004	0.952 ± 0.019	0.896 ± 0.020	12.0 ± 2.3
False positives	0.4	Full	0.949 ± 0.004	0.949 ± 0.019	0.887 ± 0.022	13.0 ± 2.3
False positives	0.4	Grouped trust	0.949 ± 0.004	0.945 ± 0.022	0.875 ± 0.029	14.4 ± 3.6
False positives	0.6	Fixed trust	0.965 ± 0.003	0.959 ± 0.015	0.903 ± 0.018	11.2 ± 2.2
False positives	0.6	Full	0.965 ± 0.003	0.953 ± 0.016	0.892 ± 0.021	12.4 ± 2.3
False positives	0.6	Grouped trust	0.965 ± 0.003	0.949 ± 0.019	0.883 ± 0.027	13.4 ± 3.4

Continued on next page

Continued

Error	Nom.	Method	Realized	AUROC	F1	SHD
False positives	0.8	Fixed trust	0.982 ± 0.001	0.957 ± 0.016	0.905 ± 0.019	11.0 ± 1.9
False positives	0.8	Full	0.982 ± 0.001	0.951 ± 0.018	0.889 ± 0.025	12.6 ± 2.3
False positives	0.8	Grouped trust	0.982 ± 0.001	0.948 ± 0.020	0.882 ± 0.027	13.4 ± 3.0
No prior	—	No prior	—	0.848 ± 0.056	0.672 ± 0.062	40.2 ± 9.7

AE Threshold-Graph Stability Across Optimizer Budgets

AE.1 Standard-budget trajectories

The 75 standard-budget trajectories vary sample size, dimension, noise, and prior accuracy. Stability requires the final five recorded graphs to equal the final graph. This criterion requires agreement across distinct iterations. The counts below are computed from the full trajectories rather than relying on a stored stabilization flag. Repeated reference configurations across axes are retained and are not independent evidence for a broader population. Exactly 30 of 75 trajectories meet this final-five criterion.

Table 45: Standard-budget threshold stability by condition.

Axis	d	T	Noise	Acc.	Stable	Final edges
dimension	10	500	gaussian	0.6	4/5	9.8 ± 2.6
dimension	20	500	gaussian	0.6	2/5	38.8 ± 3.4
dimension	30	500	gaussian	0.6	0/5	82.6 ± 5.7
noise	20	500	gaussian	0.6	2/5	38.8 ± 3.4
noise	20	500	heteroscedastic	0.6	1/5	38.2 ± 3.3
noise	20	500	laplace	0.6	3/5	30.0 ± 5.2
noise	20	500	student_t	0.6	1/5	32.8 ± 3.9
prior_quality	20	500	gaussian	0.2	2/5	47.6 ± 3.6
prior_quality	20	500	gaussian	0.4	1/5	41.4 ± 3.8
prior_quality	20	500	gaussian	0.6	2/5	38.8 ± 3.4
prior_quality	20	500	gaussian	0.8	2/5	35.2 ± 1.5
prior_quality	20	500	gaussian	1.0	4/5	32.0 ± 4.1
sample_size	20	100	gaussian	0.6	0/5	64.6 ± 5.8
sample_size	20	1000	gaussian	0.6	4/5	35.6 ± 3.8
sample_size	20	500	gaussian	0.6	2/5	38.8 ± 3.4

AE.2 Extended-budget threshold diagnostics

The 60-run diagnostic suite records graph stability over three, five, and ten final iterations, acyclicity residuals, and a first-order residual proxy. Its boolean fields concern threshold graphs despite their stored "exact" naming. The residual proxy is not automatically the active-coordinate residual in Proposition 1. This suite is separate from the 60-run metric-sensitivity experiment in Appendix AA.4.

Table 46: Extended optimizer diagnostics for every condition.

Condition	Tail 3	Tail 5	Tail 10	$h(W_0)$	Gradient proxy
T_100	1/5	1/5	0/5	$2.3e-06 \pm 8.5e-07$	$2.7e-03 \pm 1.2e-04$
T_1000	5/5	4/5	2/5	$1.9e-06 \pm 0.0e+00$	$2.7e-03 \pm 2.2e-04$
acc_02	4/5	2/5	1/5	$1.9e-06 \pm 0.0e+00$	$2.6e-03 \pm 2.4e-04$
acc_04	4/5	4/5	1/5	$1.9e-06 \pm 0.0e+00$	$2.6e-03 \pm 5.1e-05$
acc_08	3/5	3/5	2/5	$1.9e-06 \pm 0.0e+00$	$2.8e-03 \pm 9.7e-05$
acc_10	5/5	4/5	4/5	$2.3e-06 \pm 8.5e-07$	$2.7e-03 \pm 1.5e-04$
d_10	5/5	5/5	5/5	$1.9e-06 \pm 0.0e+00$	$1.9e-03 \pm 1.7e-04$
d_30	2/5	1/5	0/5	$3.8e-06 \pm 0.0e+00$	$2.8e-03 \pm 7.9e-05$
noise_heteroscedastic	4/5	4/5	4/5	$1.9e-06 \pm 0.0e+00$	$2.7e-03 \pm 1.1e-04$
noise_laplace	4/5	4/5	4/5	$1.5e-06 \pm 8.5e-07$	$2.7e-03 \pm 1.1e-04$

Continued on next page

Continued

Condition	Tail 3	Tail 5	Tail 10	$h(W_0)$	Gradient proxy
noise_student_t	4/5	3/5	3/5	$1.9e-06 \pm 0.0e+00$	$2.7e-03 \pm 8.1e-05$
reference	4/5	3/5	2/5	$1.9e-06 \pm 0.0e+00$	$2.6e-03 \pm 2.0e-04$

Interpretation. Unchanged threshold graphs over a short tail do not establish exact support constancy for an optimizer neighborhood. Small acyclicity residuals measure feasibility, while the first-order proxy measures another numerical property. The three measures are complementary and neither global convergence nor structural identification follows from them alone.