PCDM: SHARED STRUCTURE AND LOCAL MODULATION FOR DYNAMIC CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Two common approaches to dynamic causal discovery model nonstationarity as smooth temporal variation or latent regime switching. One absorbs changes in mechanisms into smooth drift, blurring abrupt regime shifts into gradual variation when regimes are short. Another introduces latent states with regime-specific edges and mechanisms, allowing ordinary fluctuations to harden into pseudo-mechanisms exactly where evidence is weakest. We therefore propose the Piecewise Causal Dynamic Model (PCDM). PCDM couples state-specific effective graphs through shared candidate scores and sparse local gates. Finite-sample analysis of a restricted regression analogue characterizes conditions for sharing benefits. Shared scores alone do not ensure that local gates select useful parents. We introduce observation-informed gate restart (INFO): a posterior-weighted additive fit ranks candidate inputs, and the resulting order reassigns existing gate strengths before neural learning resumes. On calibrated additive and core–private interaction mechanisms, INFO improves state-specific edge localization and lowers finite-response error relative to matched continuation and random-reassignment controls. Matched screening controls support the value of observation-specific state information for parent selection on additive mechanisms. Additional structured-state diagnostics show that state-inference bottlenecks can be substantially reduced, while the added benefit of sharing for private-edge localization remains to be established.

1 INTRODUCTION

Causal structure is rarely stationary. Markets shift across cycles, neural activity reorganizes across task phases, and gene-regulatory programs activate different pathways under changing conditions. Static causal models average over this variation and can turn changes in mechanisms into spurious structure (Huang et al., 2020). In dynamic causal discovery, however, the central difficulty is not simply that mechanisms change (Assaad et al., 2022; Runge et al., 2019). Under short nonstationary regimes, ordinary drift can be misread as new structure, and state labels can rewrite causality.

Dynamic causal discovery methods exhibit two recurring failure modes under short regimes (Peters et al., 2016; Assaad et al., 2022). One rewrites switching as drift by treating nonstationarity as smooth temporal variation (Primiceri, 2005; Song et al., 2009). The other introduces latent regimes or explicit state switching, with some formulations letting each state carry its own graph and parameter family (Saggioro et al., 2020; Rahmani & Frossard, 2025). Once states are unobserved, each additional state reopens support decisions over edges and mechanisms, making it easier for ordinary fluctuations to harden into pseudo-mechanisms. Under short regimes, the decisive question is therefore not whether a model can switch, but whether its state labels are allowed to rewrite causality.

We start from one principle: a regime switch should not reset causal discovery. PCDM couples state-specific effective graphs through shared candidate scores and sparse local gates.

A second difficulty is selecting the right inputs. Shared scores reduce repeated parameter estimation, but each state's gates still have to choose which candidate parents enter its mechanism. Reconstruction-based learning can settle on inputs that explain observations without recovering the active support. We address this step with observation-informed gate restart (INFO). The current neural posterior weights an additive fit to training transitions. Component magnitudes rank candi-

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

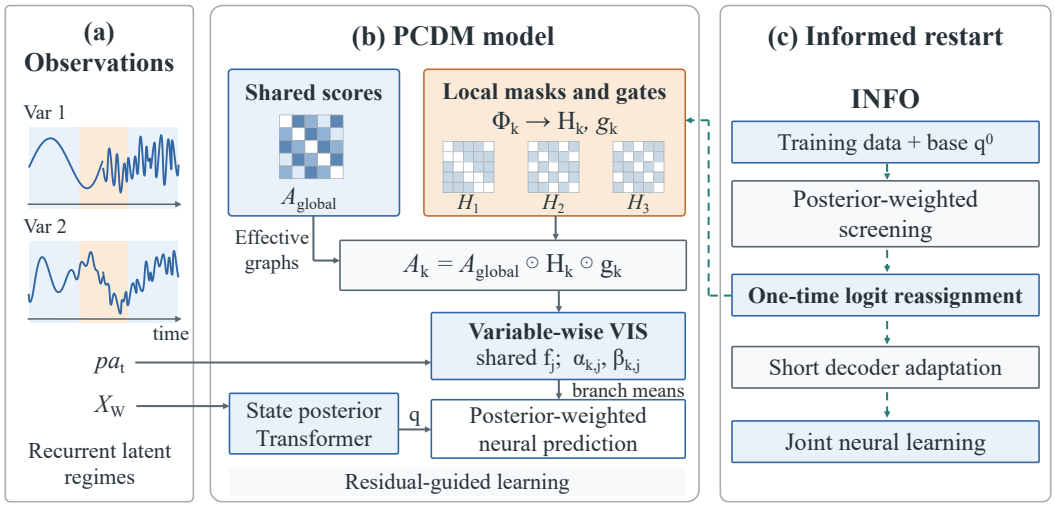

Figure 1: PCDM with observation-informed gate restart. (a) Observations contain recurrent latent regimes. (b) Shared scores, state-specific masks and gates define the effective graphs passed to variable-wise neural mechanisms. The offline posterior weights their predictions. (c) INFO uses posterior-weighted observations to reassign existing gate logits once, followed by decoder adaptation and joint learning. Solid arrows show model computations; dashed arrows show the restart path.

date inputs and reassign the existing gate values before neural learning resumes. The final predictor remains the shared neural model.

A third difficulty is benchmark comparability. Closely related methods often return different output objects. If the output object, export rule, and projection rule are not fixed in advance, comparison can drift with thresholding, repair, and projection choices rather than expose differences in structure learning itself (Assaad et al., 2022). We therefore treat the fixed export protocol as part of the comparison contract, not as a downstream implementation detail.

We study offline structural discovery under recurrent latent regimes, focusing on mechanisms that reuse stable support with sparse state-specific changes. Support coverage, state identity and local-edge activation are separate evaluation targets. The main contributions are threefold:

- At the modeling level, we couple state-specific effective graphs through shared candidate scores and sparse local gates. A finite-sample analysis in a restricted regression analogue quantifies support-search savings and the bias introduced by local departures.
- At the learning level, we introduce posterior-weighted observation screening and one-time gate reassignment, followed by decoder adaptation and joint neural learning.
- At the evaluation level, matched-budget restarts show gains in complete-candidate and state-specific edge recovery on additive and interacting mechanisms. Screening-weight controls test the role of observation-specific state information, while known-state, fixed-support and conditional-function comparisons distinguish parent selection, state inference and subsequent fitting.

2 Related Work

2.1 Dynamic Causal Discovery Under Nonstationarity

Within dynamic causal discovery, two common approaches to nonstationarity have distinct failure modes under short regimes. The first treats nonstationarity as smooth temporal variation through temporal smoothness, local stationarity, or slowly varying parameters (Primiceri, 2005; Song et al., 2009). Related temporal-dependence methods, not all of which assume smooth change, include PCMCI (Runge et al., 2019; Runge, 2020), neural Granger approaches (Granger, 1969; Tank et al., 2022; Khanna & Tan, 2019; Marcinkevičs & Vogt, 2021), DYNOTEARS (Pamfil et al., 2020), and

NTS-NOTEARS (Sun et al., 2023). RHINO models nonlinear temporal relationships with instantaneous effects and history-dependent noise (Gong et al., 2022). For formulations that impose temporal smoothness, abrupt transitions and short regimes can create an assumption mismatch: strong temporal regularization may blur boundaries. The second line introduces latent regimes or explicit state switching in unobserved-state settings (Murphy, 2002; Friedman et al., 2013; Grzegorczyk & Husmeier, 2011; Saggioro et al., 2020). Its advantage is greater flexibility in representing regime variation; in our setting, however, the main difficulty is not expressivity itself, but repeated support adjudication. Once each state carries its own graph and parameter family, the question of which edges may exist is reopened state by state. Under short regimes, this enlarged decision space makes it easier for ordinary fluctuations to be organized into apparently coherent pseudo-mechanisms rather than genuine changes in mechanisms. PCDM preserves regime variation while coupling state-specific effective graphs through shared candidate scores and sparse local gates. These methods return coefficients, scores, test statistics or graphs, making the evaluated structural object an essential part of benchmark comparison. Beyond structural modeling, inferred latent states also require mechanism-relevant semantics.

2.2 Nonlinear Screening for Structural Learning

Nonlinear screening can narrow the candidate input set before fitting a more flexible predictor. Sparse additive models (SpAM) select component functions through a sparsity penalty and sparse backfitting (Ravikumar et al., 2009). Nonparametric independence screening (NIS) fits each candidate separately and ranks its marginal function strength before joint modeling (Fan et al., 2011). Causal additive models (CAM) combine preliminary neighborhood selection, variable-order search, and sparse regression in an additive DAG model with observed variables and independent errors (Bühlmann et al., 2014). PCDM uses additive fitting to guide input selection within inferred recurrent states. The base model's fixed posterior weights a joint ridge fit over candidate lagged inputs, and the root-mean-square magnitude of each centered component determines its rank. INFO then reassigns the existing gate logits once, preserving their values and the parent budget, before neural learning resumes. The screening fit therefore supplies a state-specific starting support for the neural decoder; the final predictor remains neural. The distinction lies in connecting posterior-weighted candidate assessment to a gate restart, rather than estimating a sparse additive predictor or a variable order as the final structural model.

2.3 Shared and Hierarchical Structural Learning

Structural reuse also underlies multi-task and hierarchical learning. Inductive transfer couples Bayesian-network structures across related tasks (Niculescu-Mizil & Caruana, 2012), and joint multiple-DAG estimation allows related units to have different graphs (Oates et al., 2016). Non-homogeneous dynamic Bayesian networks already combine latent changepoints with a shared graph and segment-specific parameters; hierarchical coupling further shares information between segments (Grzegorczyk & Husmeier, 2011; 2013). These formulations establish the connection between structural reuse and latent segmentation.

PCDM binds shared candidate scores and state-local activation within short recurrent regimes, with observation-informed gate initialization. Table 1 compares the resulting structural choices. Support-union recovery and shared-plus-sparse multi-task models already establish statistical benefits under suitable overlap assumptions (Obozinski et al., 2011; Jalali et al., 2010). Our analysis in Appendix A makes those requirements explicit for a restricted regression analogue, and the structural comparisons in Section 4.1 test the role of local modulation.

3 Methodology

PCDM models recurrent regimes through a shared causal basis and sparse local modulation. Detailed proofs are provided in Appendix A.

Table 1: Structural choices in the cited neighboring formulations.

Formulation	Context	Structural reuse	Local variation	Fitting signal
Non-homogeneous DBN	Latent changepoints	Shared graph	Coupled segment parameters	Bayesian likelihood
Multiple DAGs	Related observed units	Coupled graph prior	Unit-specific edges	Given unit identities
Regime-PCMCI	Latent regimes	Regime-specific parents	Regime coefficients	Residual fit; persistence
PCDM	Latent recurrent regimes	Shared score	Masks, gates and affine terms	Residual responsibilities; input screening

3.1 Problem Setup and Structural Constraint

Consider $X_{1:T} = \{x_t\}_{t=1}^{T}$, with $x_t \in \mathbb{R}^d$, and a latent state $S_t \in \{1, \dots, K\}$(Rabiner, 1989; Murphy, 2002). Parents are restricted to earlier time steps, so the time-unrolled graph is acyclic (Runge, 2020; Pamfil et al., 2020; Assaad et al., 2022).

PCDM assigns each candidate edge a shared score and lets its strength vary across states. A per-state top-k mask limits the number of parents for each target. Training learns the scores and mechanisms together. We evaluate the resulting effective graphs in two ways: $U_{ij} = \max_k (A_k)_{ij}$ ranks whether an edge is present in any state, while the individual $(A_k)_{ij}$ scores indicate which state uses it. The union score is unchanged by refactorizations that preserve the effective graphs (Proposition A.7).

3.2 Decoupling Generation and Inference

The generative layer predicts observations; the inference layer estimates state probabilities (Kingma & Welling, 2013).

1) State-conditioned modulation at the generative layer: The latent state modulates the shared causal basis:

$$x_t = \mathcal{M}\big(PA(x_t); A_{\text{global}}, S_t, \theta_{\text{base}}\big) + \varepsilon_t, \tag{1}$$

Here $PA(x_t)$ contains lagged parent features, A_{global} is the shared basis, and θ_{base} contains shared mechanism parameters. States control gates and affine offsets.

2) Offline posterior inference: A Transformer-based (Vaswani et al., 2017) amortized inference network estimates $q_\phi(S_t \mid X_W)$ from the observation window X_W, using past and future context. Its K logits are converted to posterior probabilities by softmax:

$$\pi_t = (\pi_{t,1}, \dots, \pi_{t,K}), \qquad \pi_{t,k} = q_\phi(S_t = k \mid X_W). \tag{2}$$

3.3 Shared Causal Basis, Sparse Gates, and VIS Dynamics

For each state k, a parameter matrix $\Phi_k \in \mathbb{R}^{d \times d}$ defines a temperature-controlled gate over the shared basis:

$$g_k = \sigma\left(\frac{\Phi_k}{\tau}\right), \tag{3}$$

where $\tau > 0$ is the temperature parameter. The effective graph is

$$A_k = A_{\text{global}} \odot G_k, \qquad G_k = H_k \odot g_k, \tag{4}$$

where $\odot$ denotes element-wise multiplication. A_{global} supplies shared candidate scores, and g_k controls their continuous expression in state k.

The binary column-wise top-k mask H_k ranks gate logits and limits each target's incoming edges. A_k is the effective graph actually passed to the decoder. A *straight-through estimator* (STE) supplies gate gradients. Sharing couples the continuous parameters; hard supports are selected separately for each state.

Variable-wise input-selective (VIS) dynamics use one local mechanism per target variable, receiving only parent features selected by the structural mask. Under state k, the conditional mean of variable j is

$$\mu_{t,j}^{(k)} = \alpha_{k,j}\, f_{\text{base},j}\big(pa_{t,j} \odot A_k[:, j]\big) + \beta_{k,j}, \tag{5}$$

Here $pa_{t,j}$ contains lagged parent features, $A_k[:, j]$ gives the incoming edge scores, and $f_{\text{base},j}$ is shared across states. The scale $\alpha_{k,j}$ and offset $\beta_{k,j}$ allow state-specific changes.

Posterior-weighted means give the reconstruction, $\widehat{x}_t = \sum_k \pi_{t,k}\mu_t^{(k)}$.

3.4 Residual-Guided Fitting and Posterior Learning

Training alternates between estimating state responsibilities and fitting the neural model. A forward–backward calculation uses each branch's reconstruction error to obtain $r_{t,k}$, the probability assigned to state k. These responsibilities are held fixed while the decoder fits observations and the posterior learns to predict the responsibilities. For valid window–target exposures $\mathcal{V}$, the two base objectives are

$$\begin{aligned} L_\theta &= \frac{1}{|\mathcal{V}|d} \sum_{t\in\mathcal{V},k} \text{stopgrad}(r_{t,k}) \|x_t - \mu_t^{(k)}\|_2^2 + R_{\text{structure}}, \\ L_\phi &= -\frac{1}{|\mathcal{V}|} \sum_{t\in\mathcal{V},k} \text{stopgrad}(r_{t,k}) \log \pi_{t,k}, \\ R_{\text{structure}} &= \lambda_{\text{shared}} L_{\text{shared}} + \lambda_{\text{sparse}} L_{\text{sparse}}, \\ L_{\text{shared}} &= \frac{1}{Kd} \sum_{k,j} \big[(\alpha_{k,j} - 1)^2 + \beta_{k,j}^2\big]. \end{aligned} \tag{6}$$

The sparsity term penalizes effective edge scores, while L_{shared} keeps state-specific scales and offsets near the shared mechanism. An initial adaptation stage fits the decoder using fixed responsibilities. Subsequent updates alternate between the decoder and structure parameters θ and the posterior parameters ϕ.

3.5 Observation-Informed Gate Restart

INFO uses observation-based source rankings to reassign existing gate strengths before neural learning resumes (Figure 2). It targets parent assignment; matched-support spline references in Section 4.2 assess the remaining neural fitting error. Let u_t contain standardized lagged inputs. For each state k and target j, the selected base posterior q^0 weights a cubic-spline ridge fit on training transitions:

$$\min_{b,\{\beta_i\}_{i\neq j}} \frac{\sum_t q_{t,k}^0 \big[x_{t,j} - b - \sum_{i\neq j} \psi_i(u_{t,i})^\top \beta_i\big]^2}{\sum_t q_{t,k}^0} + \lambda_{\text{fit}} \sum_{i\neq j} \|\beta_i\|_2^2. \tag{7}$$

The bases ψ_i use training data only. A common penalty for all state–target fits is selected by observation-validation reconstruction. This uses all $d-1$ non-self candidates; the neural parent budget is imposed after ranking. These conditional observational scores guide initialization and need not identify causal parents when proxies or interaction effects dominate.

For state k, source i and target j, the component score is

$$s_{kij} = \left[\frac{1}{N}\sum_{n=1}^{N} \big(h_{kij}(u_{n,i}) - \bar{h}_{kij}\big)^2\right]^{1/2}, \qquad \bar{h}_{kij} = \frac{1}{N}\sum_n h_{kij}(u_{n,i}). \tag{8}$$

Here $h_{kij}(u) = \psi_i(u)^\top \widehat{\beta}_{kij}$, and u_n ranges over the training-parent marginal. Algorithm 1 assigns the largest existing non-self gate logit to the highest-ranked source, and continues in rank order. This preserves each column's logit multiset, temperature and parent budget. Shared scores, mechanisms and the posterior retain their selected base values.

Both Adam states are reset. The decoder and state-affine terms first adapt with the posterior and structure held fixed, then residual-guided learning updates the neural model.

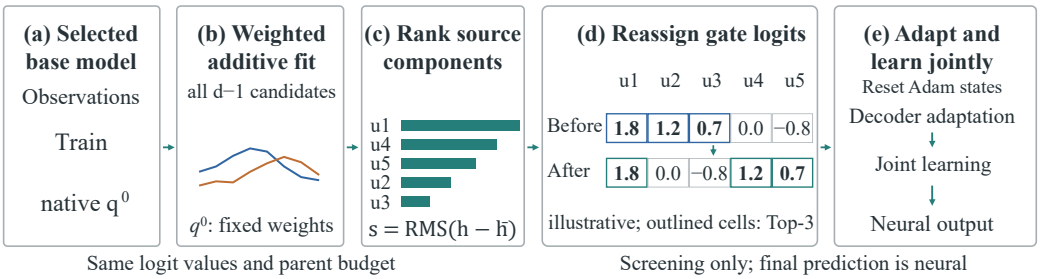

Figure 2: Observation-informed gate restart. (a) The selected base model supplies native state probabilities. (b) A posterior-weighted additive fit uses the candidate inputs. (c) Centered component magnitudes rank their contributions. (d) Existing gate logits are reassigned by this ranking, preserving their values and the parent budget. (e) Optimizer reset and decoder adaptation precede joint neural learning. The five-input Top-3 example is illustrative; the final predictor is neural.

Observation-validation reconstruction selects the returned checkpoint, and all predictions use the neural decoder. Basis construction, tie rules and the complete base-plus-restart budget are given in Appendix D.

Algorithm 1 Observation-informed gate restart (INFO)

Require: Training and validation observations $D_{\mathrm{tr}}, D_{\mathrm{val}}$; selected base model Θ^0

$q^0 \leftarrow q_{\Theta^0}(D_{\mathrm{tr}})$; fit additive components h using weights q^0
Select the additive fit by reconstruction on D_{val}; compute s by Eq. (8)
$\Theta \leftarrow \Theta^0$
for all states k and targets j **do**
  $\pi \leftarrow \mathrm{argsort}^{\downarrow}_{i \neq j} s_{kij}$; $a \leftarrow \mathrm{sort}^{\downarrow}(\Phi^0_{k,-j,j})$
  Assign $\Phi_{k,\pi(r),j} \leftarrow a_r$ for every rank r
end for
Reset optimizer states; initialize candidate set $\mathcal{C} \leftarrow \{\Theta\}$
Adapt decoder and affine terms using q^0, holding structure and posterior fixed; append iterates to $\mathcal{C}$
while joint-learning budget remains **do**
  Infer responsibilities; update decoder, structure and posterior using Eq. (6)
  Update temporal parameters; append Θ to $\mathcal{C}$
end while
return Earliest $\Theta \in \mathcal{C}$ minimizing neural-posterior reconstruction on D_{val}

4 Experiments

We examine state-specific structure, informed parent selection, and the information and learning stages that support it. Controlled comparisons use $d = 20$ and three parents per state and target. Each repeated unit is a graph/data/training tuple; response errors average three fresh observations within each graph before computing means and sample SDs across five tuples. Observation-validation reconstruction selects the earliest minimum checkpoint, followed by one fixed validation state permutation. We report ordinary paired 95% t intervals and Holm-adjusted p values for prespecified comparisons.

Appendix B supplies external-baseline comparisons, public-benchmark protocols and auxiliary-loss comparisons; Appendix C details controlled data and state diagnostics. Appendices D–E provide informed-support controls, full outcomes and supplementary ablations.

4.1 Shared Structure and State-Specific Edge Recovery

We compare Single-graph, Shared-only, PCDM and Independent structures. Shared-only retains regime-dependent mechanism terms; Independent learns separate state supports with the same decoder architecture. At the central tanh support-rewiring condition, Single-graph leads union ranking.

An edge must be assigned to its active state. The additive-sine experiment gives each of 20 targets two stable parents and a distinct third parent in each of three recurring states. Finite-perturbation calibration checks that all 180 active coordinates exceed 1.5 noise SD. PRIVATE AP pools 60 active and 120 inactive edge–state coordinates from the 60 private edges, using the decoder's effective graphs. Owner accuracy divides credit equally among exact maximum-score ties over all 60 edges. Signed margin subtracts the largest wrong-state score from the true-state score.

With given states, PCDM attains 53.89% owner accuracy, compared with one-third tie credit for Single-graph and Shared-only (Table 2). Only 36% of its private edges have a strictly positive margin. All three localization intervals relative to Independent include zero. Correct state information helps recovery, but these comparisons do not establish an additional sharing advantage.

Table 2: Held-out state information and private-edge localization. Mean ± sample SD; owner is in percent. Both structures use the same learning protocol within each block.

State protocol	Structure	PRIVATE AP ↑	Owner (%) ↑	Margin ↑	State ARI ↑
Given states	PCDM	$\mathbf{0.552 \pm 0.024}$	$\mathbf{53.89 \pm 3.02}$	$\mathbf{0.108 \pm 0.020}$	given
	Independent	0.536 ± 0.032	51.33 ± 3.16	0.081 ± 0.027	given
Mixture reconstruction	PCDM	0.353 ± 0.012	$\mathbf{35.67 \pm 3.69}$	$\mathbf{-0.024 \pm 0.013}$	0.001 ± 0.003
	Independent	$\mathbf{0.361 \pm 0.012}$	35.33 ± 1.91	-0.026 ± 0.008	$\mathbf{0.005 \pm 0.008}$
Residual-guided	PCDM	$\mathbf{0.508 \pm 0.051}$	48.78 ± 7.44	$\mathbf{0.074 \pm 0.034}$	$\mathbf{0.651 \pm 0.253}$
	Independent	0.505 ± 0.061	$\mathbf{49.22 \pm 5.85}$	0.063 ± 0.037	0.541 ± 0.240
Structured diagnostic	PCDM	0.689 ± 0.041	67.64 ± 3.70	0.183 ± 0.020	$\mathbf{0.945 \pm 0.022}$
	Independent	$\mathbf{0.707 \pm 0.021}$	$\mathbf{68.19 \pm 1.84}$	$\mathbf{0.186 \pm 0.011}$	0.928 ± 0.060

The first three blocks use five tuples. The structured diagnostic reports the four non-development tuples under the fixed protocol; all five are in Appendix C.

State recovery improves with residual-guided learning and structured-state inference (Table 2), yet the structured diagnostic uniquely locates only 33 of 60 private edges, close to Independent.

Candidate-score tying. With identical INFO initialization, fixed posterior and shared VIS, tied-minus-untied response NMSE at $L = 64$ is -0.01081 (paired 95% CI $[-0.01675, -0.00486]$; six-test Holm $p = 0.0435$). PRIVATE AP, the $L = 32$ response difference and both length interactions show no clear change.

4.2 Observation-Informed Learning and Mechanism Fitting

Matched restart controls. We compare INFO with KEEP and RANDOM. KEEP retains the base gate assignment; RANDOM reassigns the same values while matching INFO's initial-support overlap in each state–target column. All three use the same restart budget, VIS, unit input gain and spectral normalization. We analyze five additive and five interaction tuples separately. ALL-STATE AP ranks all 1,140 directed nonself edge–state coordinates (180 active and 960 inactive), using the selected decoder's effective scores.

Finite-response evaluation. Let $\widehat{m}_j(u; q_t)$ denote the neural mean at lagged input u with the factual posterior q_t fixed. Only the lagged source coordinate is perturbed:

$$\widehat{R}_{tij} = \widehat{m}_j(u_t + \delta_i e_i; q_t) - \widehat{m}_j(u_t; q_t), \qquad \mathrm{NMSE} = \frac{\sum_{t,i,j}(\widehat{R}_{tij} - R^*_{tij})^2}{\sum_{t,i,j}(R^*_{tij})^2}. \tag{9}$$

Here R^* is evaluated from the controlled generating mechanism after training, with neither its labels nor its function used for fitting or selection. Responses are compared in training-standardized output coordinates. Physical perturbations are $\sigma_i \delta_i$; physical responses are divided by training SD σ_j. An

identically zero predicted response gives NMSE one. Perturbation signs, source spans and target sampling are fixed across the compared models.

Table 3: Observation-informed learning across additive and interaction mechanisms. Mean ± sample SD over five graph–data–training tuples per family.

Model	ALL-STATE AP ↑	PRIVATE AP ↑	Response NMSE ↓	Correct/60 ↑
Additive mechanisms				
KEEP	0.469 ± 0.009	0.543 ± 0.023	0.922 ± 0.010	20.8 ± 2.9
RANDOM	0.493 ± 0.030	0.537 ± 0.022	0.927 ± 0.016	20.2 ± 2.2
INFO	$\mathbf{0.981 \pm 0.020}$	$\mathbf{0.975 \pm 0.030}$	$\mathbf{0.764 \pm 0.012}$	$\mathbf{58.0 \pm 2.8}$
Core–private interactions				
KEEP	0.547 ± 0.029	0.584 ± 0.070	0.903 ± 0.022	25.0 ± 4.8
RANDOM	0.519 ± 0.053	0.535 ± 0.058	0.919 ± 0.020	21.2 ± 4.1
INFO	$\mathbf{0.922 \pm 0.048}$	$\mathbf{0.939 \pm 0.058}$	$\mathbf{0.823 \pm 0.032}$	$\mathbf{55.0 \pm 4.4}$

Correct requires a unique correct owner among 60 private edges; fair owner accuracy is reported in Table 16. Responses average three fresh trajectories within each tuple. Bold marks the best neural mean per family. Structural columns use actual neural effective-edge scores.

On additive mechanisms, INFO improves PRIVATE AP from KEEP's 0.543 to 0.975 and lowers response NMSE from 0.922 to 0.764. The paired changes are $+0.432$ $[0.412, 0.452]$ and -0.158 $[-0.170, -0.145]$. Both endpoints improve on every graph against both controls.

Core–private interactions. INFO also improves recovery on mechanisms with a product interaction between a stable and a state-private parent. It uniquely locates 55 of 60 private edges on average, versus 25 for KEEP and 21.2 for RANDOM. Relative to KEEP, AP increases by 0.355 $[0.291, 0.418]$ and response NMSE decreases by 0.080 $[0.049, 0.111]$. Relative to RANDOM, the changes are $+0.404$ $[0.329, 0.478]$ and -0.096 $[-0.131, -0.062]$. Each mechanism family has four primary tests against KEEP and RANDOM; all eight pass their family-wise Holm correction.

4.3 Contributions of Screening Information and Subsequent Learning

Posterior information in screening. We keep the base model, spline basis, ridge grid, gate values and neural learning budget fixed. NATIVE uses the INFO base posterior for each training observation. POOLED replaces it with the training-average state weights; PERMUTED shuffles whole posterior vectors within each independent training trajectory. Only the screening fit uses these altered weights. Subsequent neural learning follows the common protocol.

Table 4: Posterior weighting used for support screening. The neural models retain the same training and selection protocol.

Mechanisms	Screening	PRIVATE AP ↑	Response NMSE ↓	Correct/60 ↑
Additive	NATIVE	$\mathbf{0.975 \pm 0.030}$	$\mathbf{0.764 \pm 0.012}$	**58.0**
	POOLED	0.779 ± 0.130	0.803 ± 0.021	42.2
	PERMUTED	0.726 ± 0.119	0.806 ± 0.016	38.8
Interactions	NATIVE	$\mathbf{0.939 \pm 0.058}$	$\mathbf{0.823 \pm 0.032}$	**55.0**
	POOLED	0.790 ± 0.125	0.838 ± 0.031	44.0
	PERMUTED	0.825 ± 0.093	0.833 ± 0.033	47.0

AP and NMSE are mean ± sample SD over five paired graph/data/training instances per family. Three fresh response trajectories are averaged within each graph. Correct/60 is the mean strict correct-owner count. Bold marks the best mean within each family and column.

In the additive family, native weighting improves PRIVATE AP over pooled and permuted screening by 0.195 and 0.249, and lowers response NMSE by 0.038 and 0.042; all four prespecified within-family Holm tests pass. In the interaction family, the point estimates favor native weighting, but none of the four contrasts passes the same correction.

Learning after informed assignment. High complete-candidate ranking quality is already present immediately after reassignment: ALL-STATE AP is 0.967 on additive mechanisms and 0.947 on interaction mechanisms. Subsequent learning changes these values to 0.981 and 0.922, respectively; on the interaction family, true hard-support coordinates decrease from 861 to 829 across the five graphs (Tables 23 and 24).

Fixing the informed binary support throughout neural learning gives additive PRIVATE AP 0.947 and response NMSE 0.765, compared with INFO's 0.975 and 0.764. The paired intervals for the effect of support changes include zero. Both models receive the full neural learning budget.

At the same three-parent budget, additive spline references reach response NMSE 0.110 on the screening support and 0.040 on INFO's selected support and posterior. On interaction mechanisms, the latter comparison gives spline-minus-neural NMSE -0.622 $[-0.695, -0.548]$. A substantial neural function-fitting gap remains after parent selection improves.

Public-network recovery. On DREAM3, INFO raises macro AUROC/AP from KEEP's 52.14%/3.28% to 53.92%/3.75%; the paired gains of 1.77 and 0.47 percentage points pass the two-endpoint Holm correction. Inference pairs training seeds after averaging the five fixed networks. NTS-NOTEARS attains the highest macro AUROC (56.44%), whereas DYNOTEARS attains the highest macro AP (7.39%); both exceed INFO on both macro metrics. External baselines, detailed internal controls and supplementary dynamic/static experiments at low and high dimensions are reported in Appendix B.

5 Conclusion

PCDM represents recurrent mechanism changes as sparse local modulation over a shared causal basis. This formulation links a common structural representation to state-specific effective graphs. INFO uses posterior-weighted observations to guide their initial parent assignments. The finite-sample analysis identifies conditions under which sharing reduces estimation error; the controlled evaluation measures whether recovered edges are assigned to the states that use them.

Across additive mechanisms and mechanisms with core–private interactions, INFO improves private-edge localization and reduces finite-response error relative to both matched restart controls. These results support the central contribution of this work: posterior-weighted observations can guide state-specific parent assignment within a shared neural model, while preserving the parent budget and retaining a neural predictor. The remaining response-error gap to matched-input spline references motivates further work on neural mechanism fitting.

Reproducibility Statement

The appendices provide the theory, data-generation and admission rules, model settings, seeds, selection protocols and per-graph comparisons. Restart and interaction experiments retain every pre-specified graph and both matched controls. The supplement provides executable implementations and bundled datasets, with separate commands for controlled-experiment training, technical checks, and reaggregation of the published results. Its README states the dependencies and inputs.

AI Use Statement

The authors led the scientific conception, core architecture, experimental design, manuscript development and figure design. Under their direction, ChatGPT (GPT-5.6 and GPT-6) assisted code implementation and debugging, result checking, manuscript drafting and language polishing, and preparation of controlled experimental datasets. The authors reviewed the resulting code, analyses, text and figures and take full responsibility for the work.

REFERENCES

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

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. doi: 10.1214/14-AOS1260. URL https://doi.org/10.1214/14-AOS1260.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. doi: 10.1198/jasa.2011.tm09779. URL https://doi.org/10.1198/jasa.2011.tm09779.

Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. *arXiv preprint arXiv:1301.7374*, 2013.

Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *arXiv preprint arXiv:2210.14706*, 2022. doi: 10.48550/arXiv.2210.14706. URL https://arxiv.org/abs/2210.14706.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

Marco Grzegorczyk and Dirk Husmeier. Non-homogeneous dynamic bayesian networks for continuous data. *Machine Learning*, 83(3):355–419, 2011.

Marco Grzegorczyk and Dirk Husmeier. Regularization of non-homogeneous dynamic bayesian networks with global information-coupling based on hierarchical bayesian models. *Machine Learning*, 91:105–154, 2013. doi: 10.1007/s10994-012-5326-3. URL https://link.springer.com/article/10.1007/s10994-012-5326-3.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

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.

Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K. Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 23, pp. 964–972, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/hash/00e26af6ac3b1c1c49d7c3d79c60d000-Abstract.html.

Saurabh Khanna and Vincent YF Tan. Economy statistical recurrent units for inferring nonlinear granger causality. *arXiv preprint arXiv:1911.09879*, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ričards Marcinkevičs and Julia E. Vogt. Interpretable models for granger causality using self-explaining neural networks. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.07600. Author manuscript, arXiv:2101.07600.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.05957.

Kevin P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, EECS Department, University of California, Berkeley, 2002. URL https://www2.eecs.berkeley.edu/Pubs/TechRpts/2002/8174.html.

Alexandru Niculescu-Mizil and Rich Caruana. Inductive transfer for Bayesian network structure learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pp. 167–180. PMLR, 2012. URL https://proceedings.mlr.press/v27/niculescu12a.html.

Chris J. Oates, Jim Q. Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26:797–811, 2016. doi: 10.1007/s11222-015-9570-9. URL https://link.springer.com/article/10.1007/s11222-015-9570-9.

Guillaume Obozinski, Martin J. Wainwright, and Michael I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011. doi: 10.1214/09-AOS776. URL https://arxiv.org/abs/0808.0711.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pp. 1595–1605. Pmlr, 2020.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Jean Ponce, Basile Terver, Martial Hebert, and Michael Arbel. Dual perspectives on non-contrastive self-supervised learning, 2026. Version 3, 5 February 2026.

Giorgio E Primiceri. Time varying structural vector autoregressions and monetary policy. *The Review of economic studies*, 72(3):821–852, 2005.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4546–4554. PMLR, 2025. URL https://proceedings.mlr.press/v258/rahmani25a.html.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5):1009–1030, 2009. doi: 10.1111/j.1467-9868.2009.00718.x. URL https://academic.oup.com/jrsssb/article/71/5/1009/7092930.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on uncertainty in artificial intelligence*, pp. 1388–1397. Pmlr, 2020.

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

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://pubmed.ncbi.nlm.nih.gov/15845847/.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11):113115, 2020. doi: 10.1063/5.0020538. URL https://arxiv.org/abs/2007.00267.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011. doi: 10.1016/j.neuroimage.2010.08.063. URL https://www.sciencedirect.com/science/article/pii/S1053811910011602.

Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. *Advances in neural information processing systems*, 22, 2009.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. NTS-NOTEARS: Learning nonparametric DBNs with prior knowledge. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1942–1964. PMLR, 2023. URL https://proceedings.mlr.press/v206/sun23c.html.

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.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7:43, 2006. doi: 10.1186/1471-2105-7-43. URL https://link.springer.com/article/10.1186/1471-2105-7-43.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pp. 11–30. Springer, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. *Advances in Neural Information Processing Systems*, 37:4315–4340, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A Detailed Theoretical Analysis

The analysis develops support-selection, finite-sample sharing, state-misassignment and response-consistency results.

A.1 Identification Targets and the Shared-Support Setting

In the lag-only setting, an edge $(i, t-\ell) \to (j, t)$ has $\ell > 0$, so temporal order makes the time-unrolled graph acyclic (Peters et al., 2017; Assaad et al., 2022). We distinguish shared candidate scores from effective state structures.

State-label permutations and product-preserving transformations can preserve the effective structures. In controlled validation, CORE is the intersection of true state supports, LOCAL is their union minus CORE, and ABSENT is the complement. Shared logits score candidates, while state masks determine effective sparsity.

Sparse local modulation permits activation, suppression and limited parameter variation over reusable support. Table 5 lists the working conditions and their diagnostics.

Table 5: Working conditions A0–A4 and corresponding diagnostics.

	Required condition	Relevant evidence
A0	Locally comparable mechanisms within the current window	Valid pair rate; fragmentation; within-state response variation
A1	No dominant affine, posterior or other non-structural bypass	Matched structural-path and shortcut controls
A2	Mechanisms differ along reachable perturbation directions	Between-mechanism response separation; direction coverage
A3	Responses remain nondegenerate	Response magnitude relative to the normalization floor
A4	The structural bottleneck is effective	Per-state budget; density; false positives; exported-object agreement

A.2 Support Selection, Reconstruction, and Response Conditions

A.2.1 Support-Adjudication Compression

Proposition A.1 (Counting effective support families)**.** *For M candidate edges and K states, unrestricted binary state supports give 2^{KM} patterns. If the effective union has exactly s edges, the number is instead $\binom{M}{s}(2^K-1)^s$. Selecting a common candidate set of size s without requiring every candidate to occur gives at most $\binom{M}{s}2^{Ks}$ descriptions. A once-selected candidate set therefore leaves state-specific activation decisions.*

Proof. Choose the s union edges and assign each a nonempty subset of the K states. There are $2^K - 1$ such subsets per edge. For a candidate set rather than an exact union, the empty subset is also allowed; some descriptions may coincide. The unrestricted count independently assigns each of the KM binary entries. These counts include local activation and do not reduce the complete decision burden from KM to M. □

Appendix A.4 extends the count to unknown-support selection, local departures, noise and approximation error.

A.2.2 Reconstruction-Only Insufficiency

Proposition A.2 (A finite-data counterexample)**.** *If a model class can realize any prescribed hard assignments on a finite set of distinct inputs and can interpolate arbitrary targets within each assigned group, empirical reconstruction alone does not identify the true state partition.*

Proof. For any partition of the input set, define each group's predictor to interpolate its assigned targets, and select it with probability one on those inputs. The empirical squared loss is zero for every such partition. Two partitions that are not state permutations thus attain the same global empirical minimum. For one-dimensional distinct inputs, ordinary interpolation polynomials provide explicit group predictors. This counterexample concerns that expressive function class, not a proof that the finite-width shared tanh network realizes every partition. □

A.3 Finite-Sample Sharing in a Restricted Regression Model

We analyze when coefficient pooling reduces noise and when heterogeneity introduces bias. Unlike distribution-free uniform-convergence theory (Vapnik & Chervonenkis, 2015), this calculation uses an explicit linear estimator. Related support-union and shared/local sparsity analyses are given by Obozinski et al. (2011); Jalali et al. (2010).

Model and assumptions. There are $K \geq 2$ known regimes, each with n samples and p candidate lagged features. In regime k,

$$y_k = X_k\beta_k + \varepsilon_k, \qquad X_k^\top X_k = nI_p, \qquad \varepsilon_k \sim \mathcal{N}(0, \sigma^2 I_n). \tag{10}$$

Designs and regime labels are fixed, noises are independent across regimes, and $n \geq p$. Set

$$Z_k = n^{-1}X_k^\top y_k, \quad \bar{Z} = K^{-1}\sum_k Z_k, \quad \bar{\beta} = K^{-1}\sum_k \beta_k, \quad D^2 = K^{-1}\sum_k \|\beta_k - \bar{\beta}\|_2^2. \tag{11}$$

For a deterministic pooling weight $a \in [0, 1]$, define

$$\widehat{\beta}_k(a) = (1-a)Z_k + a\bar{Z}. \tag{12}$$

The endpoints are independent estimation and complete pooling; intermediate values retain local corrections.

Theorem A.1 (Exact finite-sample sharing–heterogeneity trade-off)**.** *Under Equation* (10)*, the average squared coefficient risk is*

$$R(a) = \frac{1}{K}\sum_k \mathbb{E}\|\widehat{\beta}_k(a) - \beta_k\|_2^2 = a^2D^2 + \frac{p\sigma^2}{n}\left[(1-a)^2 + \frac{2a-a^2}{K}\right]. \tag{13}$$

In particular,

$$R(0) = \frac{p\sigma^2}{n}, \qquad R(1) = D^2 + \frac{p\sigma^2}{Kn}. \tag{14}$$

Complete pooling has lower risk than independent estimation exactly when

$$D^2 < \frac{p\sigma^2}{n}\left(1 - \frac{1}{K}\right). \tag{15}$$

For $V = p\sigma^2/n > 0$, the risk-minimizing weight is

$$a_* = \frac{V(1-1/K)}{D^2 + V(1-1/K)}. \tag{16}$$

Proof. Write $Z_k = \beta_k + \xi_k$, where independent ξ_k have mean zero and covariance $(\sigma^2/n)I_p$. The estimation error equals

$$a(\bar{\beta} - \beta_k) + (1-a)\xi_k + a\bar{\xi}.$$

Its squared bias is $a^2\|\bar{\beta} - \beta_k\|_2^2$. The covariance is

$$\frac{\sigma^2}{n}\left[\left(1 - a + \frac{a}{K}\right)^2 + (K-1)\frac{a^2}{K^2}\right]I_p = \frac{\sigma^2}{n}\left[(1-a)^2 + \frac{2a-a^2}{K}\right]I_p.$$

The cross term between deterministic bias and mean-zero noise vanishes. Taking the trace and averaging over k proves Equation (13). Substituting $a = 0, 1$ gives the endpoint risks. Differentiating the convex quadratic gives a_*, and subtracting the two endpoint risks proves Equation (15). □

Sparse local departures and short regimes. Suppose $\beta_k = \beta_0 + \delta_k$, each δ_k has at most s nonzero coordinates, and $\|\delta_k\|_\infty \le B$. Then

$$D^2 = K^{-1} \sum_k \|\delta_k - \bar{\delta}\|_2^2 = K^{-1} \sum_k \|\delta_k\|_2^2 - \|\bar{\delta}\|_2^2 \le sB^2.$$

Thus $sB^2 < V(1 - 1/K)$ suffices for full pooling to improve risk. Smaller n increases the potential variance reduction; heterogeneity increases bias. Here n is total regime exposure, not segment length. The support-change experiment approximately fixes exposure while varying segment length, so the theorem does not predict its curve.

Corollary A.1 (A conditional pooled support bound)**.** *Let $\beta_k = \beta_0 + \delta_k$, with $b = \|\bar{\delta}\|_\infty$. For $0 < \alpha < 1$, set*

$$t_N = \sigma \sqrt{\frac{2 \log(2p/\alpha)}{Kn}}, \qquad u = b + t_N.$$

With probability at least $1 - \alpha$, $\|\bar{Z} - \beta_0\|_\infty \le u$. If every nonzero component of β_0 has magnitude strictly larger than $2u$, thresholding $\bar{Z}$ at u recovers $\mathrm{supp}(\beta_0)$.

Proof. Each component of $\bar{\xi}$ is Gaussian with variance $\sigma^2/(Kn)$. The Gaussian tail inequality and a union bound give $\Pr(\|\bar{\xi}\|_\infty > t_N) \le 2p \exp[-Knt_N^2/(2\sigma^2)] = \alpha$. Since $\bar{Z} - \beta_0 = \bar{\delta} + \bar{\xi}$, the stated bound follows. On this event, zero components have magnitude at most u, while nonzero components exceed u by the minimum-signal condition. □

The terms Kn and b represent pooled evidence and local contamination. Equating $\mathrm{supp}(\beta_0)$ with CORE requires a defined stable component and bounded contamination; shared support with coefficient cancellation is insufficient.

A.4 Unknown Support, Local Departures, and Finite-Sample Complexity

Shared edges may have opposite coefficient signs, so shared support alone does not justify coefficient pooling. Following established support-union and shared-plus-sparse analysis (Obozinski et al., 2011; Jalali et al., 2010), we instead analyze constrained least squares with unknown support and unrestricted coefficients on shared candidates.

Model and estimator. For one target variable, stack K fixed-design regressions $y_k = X_k \beta_k + \epsilon_k$, with $X_k \in \mathbb{R}^{n_k \times p}$, $N = \sum_k n_k$, and independent $\epsilon_k \sim N(0, \sigma^2 I_{n_k})$. Let $y = \mu + \epsilon$ and $X = \mathrm{diag}(X_1, \ldots, X_K)$. The design and state labels are treated as fixed throughout this analysis.

A candidate family consists of $S_0 \subseteq [p]$, $|S_0| = s_0$, and $D_k \subseteq [p] \setminus S_0$, $|D_k| \le s$, where $s_0 + s \le p$. Coefficients are unrestricted on $F_k = S_0 \cup D_k$ and zero elsewhere. Thus a shared candidate need not be active in every state, and its sign and magnitude need not be shared. S_0 is neither assumed known to the estimator nor claimed uniquely recoverable. The observable support target is the collection of effective supports $\mathrm{supp}(\beta_k)$; CORE is their intersection, not necessarily S_0. Deletions within S_0 are represented by zero coefficients, while D_k allows sparse additions.

Each family defines a prediction subspace V_F. Its rank is at most $D = K(s_0 + s)$, and the number L of distinct subspaces is bounded by

$$L \le L_{\mathrm{sh}} = \binom{p}{s_0} \left[\sum_{j=0}^{s} \binom{p - s_0}{j} \right]^K. \tag{17}$$

Repeated subspaces count once. Define $\widehat{\mu}$ by minimizing $\|y - v\|_2^2$ over their union with a fixed tie rule. The bound below applies to this constrained least-squares estimator.

Theorem A.2 (Finite-family prediction and effective-support bound)**.** *Let $0 < \alpha < 1$. Put*

$$a_L = \sqrt{2D} + \sqrt{2 \log(L/\alpha)}, \qquad b = \min_{V_F} \mathrm{dist}(\mu, V_F).$$

With probability at least $1 - \alpha$,

$$\|\widehat{\mu} - \mu\|_2 \le b + 2\sigma a_L. \tag{18}$$

Consequently, for every fixed $c > 0$, the prediction error divided by N is at most

$$(1 + c)b^2/N + 4(1 + c^{-1})\sigma^2 a_L^2/N.$$

In the well-specified case $b = 0$, the sharper bound is $4\sigma^2 a_L^2/N$.

For support recovery, additionally suppose each true effective support has exactly $m = s_0 + s$ nonzero coefficients, is in the candidate family, and $b_{\min} = \min_{k,i:\beta_{ki}\neq 0} |\beta_{ki}| > 0$. Suppose

$$\|X_k h\|_2^2 \geq n_k \kappa \|h\|_2^2 \quad \textit{for every } 2m\textit{-sparse } h \textit{ and every } k, \qquad \kappa > 0. \tag{19}$$

If $b_{\min} > 2\sigma a_L/\sqrt{n_{\min}\kappa}$, every selected effective support equals the corresponding true support on the same event. No sign agreement across states is required. Without the exact-size condition, the same argument excludes missing true edges but does not exclude extra selected coordinates or nonzero fitted false positives. If $m = 0$, the well-specified family contains only the empty support and recovery is immediate.

Proof. Choose a fixed best-approximating subspace V_* and its projection v_* of μ. They depend only on the fixed design and μ, not on noise. For each V_F, let P_F project onto $V_F + V_*$. Its rank is at most $2D$. We first establish the needed Gaussian bound. If Z is a sum of r independent squared standard normals, direct Gaussian integration gives

$$\log \mathbb{E}e^{u(Z-r)} = -ru - \tfrac{r}{2}\log(1 - 2u) \leq \frac{ru^2}{1 - 2u}, \quad 0 < u < \tfrac{1}{2}.$$

The inequality follows by expanding $-\log(1-2u)-2u$ as its nonnegative power series and bounding its coefficient of u^j by 2^{j-1} for $j \geq 2$. Chernoff's bound with $u = \sqrt{x}/(\sqrt{r} + 2\sqrt{x})$ gives $\Pr\{Z > r + 2\sqrt{rx} + 2x\} \leq e^{-x}$. Since $(\sqrt{r} + \sqrt{2x})^2 \geq r + 2\sqrt{rx} + 2x$, it follows that

$$\Pr\{\|P_F\epsilon\|_2 > \sigma(\sqrt{2D} + \sqrt{2x})\} \leq e^{-x}.$$

Rank zero is immediate. Set $x = \log(L/\alpha)$ and take a union bound over the L subspaces. With probability at least $1 - \alpha$, all these projection norms are at most $z = \sigma a_L$.

The least-squares definition implies

$$\|\widehat{\mu} - \mu\|^2 \leq \|v_* - \mu\|^2 + 2\langle \epsilon, \widehat{\mu} - v_* \rangle.$$

The difference in the inner product belongs to the sum of its selected subspace and V_*. On the simultaneous event, writing $e = \|\widehat{\mu} - \mu\|$ gives $e^2 \leq b^2 + 2z(e + b)$. If $e + b > 0$, division by $e + b$ yields $e - b \leq 2z$; the zero case is immediate. Squaring and using $4bz \leq cb^2 + 4z^2/c$ gives the stated general bound.

For the support statement, any wrong candidate with at most m coordinates must omit a true coordinate in at least one state, because the true support in that state has size m. For any coefficient vector supported on that candidate, its difference from the truth is at most $2m$-sparse and has a coordinate of magnitude at least $b_{\min}$. Equation (19) therefore places the corresponding prediction at distance at least $\sqrt{n_{\min}\kappa}\, b_{\min}$ from μ. This exceeds $2\sigma a_L$, contradicting Eq. (18) with $b = 0$. Restricted injectivity also ensures that the fitted coefficients on a correct support are unique. The statement concerns effective supports, not the nonunique decomposition into $S_0, D_1, \ldots, D_K$. □

Comparison and departures. Independent selection of at most m candidates in every state has the same rank upper bound D, but the family-count upper bound $L_{\text{ind}} = [\sum_{j=0}^{m} \binom{p}{j}]^K$. If $\log L_{\text{sh}} < \log L_{\text{ind}}$, the theorem gives a smaller certified noise envelope and a weaker sufficient minimum-signal threshold for the structured family, not a universal risk ordering. For $p = 100, K = 3, s_0 = 3, s = 1$, the natural-log bounds are 25.75 and 45.67, both with $D = 12$. The saving is in support search, not coefficient count; excluded local additions can make $b > 0$.

More explicitly, write $\beta_k = \theta_k + \delta_k$, with θ_k supported on S_0 but otherwise state-specific. Retaining a set of at most s outside-S_0 entries of δ_k gives a candidate. If r_k is the omitted vector, then $b^2 \leq \sum_k \|X_k r_k\|^2 \leq \sum_k \|X_k\|_{\text{op}}^2 \|r_k\|^2$. If all departures fit, this term vanishes. No universal rewiring percentage determines failure. Opposite signs on shared edges are permitted; $p > n_k$ is allowed only when the sparse subdesign rank conditions hold, which is not established for the observed time series.

A.5 State Misassignment and Adaptive Responsibilities

For fixed covariates $x_t \in \mathbb{R}^p$, true labels S_t, and responses $y_t = x_t^\top \beta_{S_t} + \epsilon_t$, allow responsibilities $q_{tk} \geq 0$ to depend arbitrarily on the responses. Define $m_k = \sum_t q_{tk} > 0$, $a_{tk} = q_{tk}/m_k$, $G_k = \sum_t a_{tk} x_t x_t^\top$, and the weighted least-squares estimator $\widetilde{\beta}_k = G_k^{-1} \sum_t a_{tk} x_t y_t$. The following result does not assume that the selected responsibilities are independent of noise.

Proposition A.3 (Misassignment, local mechanisms, and selection)**.** *Suppose $\|x_t\| \leq R$ and $\lambda_{\min}(G_k) \geq \kappa_k > 0$. Define the weighted confusion mass $\eta_{k\ell} = \sum_{t:S_t=\ell} a_{tk}$, $\eta_k = \sum_{\ell \neq k} \eta_{k\ell}$, and $\Delta_{k\ell} = \|\beta_\ell - \beta_k\|$. For any fixed reference probability weights a^0_{tk}, let $v_k^2 = \sum_t (a^0_{tk})^2$ and $\rho_k = \|a_{\cdot k} - a^0_{\cdot k}\|_2$. Deterministically,*

$$\|\widetilde{\beta}_k - \beta_k\| \leq \frac{1}{\kappa_k} \left[R^2 \sum_{\ell \neq k} \eta_{k\ell} \Delta_{k\ell} + \left\| \sum_t a^0_{tk} x_t \epsilon_t \right\| + R \rho_k \|\epsilon\| \right]. \tag{20}$$

If $\epsilon \sim N(0, \sigma^2 I_N)$, the labels, design and reference weights are fixed independently of ϵ, then with probability at least $1 - \alpha$, simultaneously for all k satisfying the displayed Gram condition,

$$\|\widetilde{\beta}_k - \beta_k\| \leq \frac{1}{\kappa_k} \left[R^2 \sum_{\ell \neq k} \eta_{k\ell} \Delta_{k\ell} + \sigma R v_k \sqrt{2p \log(4Kp/\alpha)} + \sigma R \rho_k \{\sqrt{N} + \sqrt{2\log(2/\alpha)}\} \right]. \tag{21}$$

The realized η, ρ, κ may be data-dependent; the probability event is uniform with respect to the choice of q.

Proof. Substitute y_t into the normal equation and subtract $G_k \beta_k$:

$$\widetilde{\beta}_k - \beta_k = G_k^{-1} \left[\sum_t a_{tk} x_t x_t^\top (\beta_{S_t} - \beta_k) + \sum_t a_{tk} x_t \epsilon_t \right].$$

The first sum has norm at most $R^2 \sum_{\ell \neq k} \eta_{k\ell} \Delta_{k\ell}$. Split the second sum into the reference-weight term and the difference term. Cauchy–Schwarz gives $\| \sum_t (a_{tk} - a^0_{tk}) x_t \epsilon_t \| \leq R \rho_k \|\epsilon\|$ without any independence assumption about q. Multiplication by $\|G_k^{-1}\| \leq 1/\kappa_k$ proves the deterministic inequality.

Each coordinate of the fixed reference sum is Gaussian with variance at most $\sigma^2 R^2 v_k^2$. A two-sided Gaussian tail bound and a union bound over Kp coordinates, allocating failure probability $\alpha/2$, bound its vector norm by the middle term of Eq. (21). The Gaussian norm inequality proved above, with rank N and failure probability $\alpha/2$, bounds $\|\epsilon\|$ by the final brace. Intersect the two events. Both involve only noise, fixed covariates and fixed reference weights, so they hold for every subsequently selected responsibility matrix. This proves the simultaneous statement without conditioning on the selected labels. □

For oracle reference weights $a^0_{tk} = \mathbf{1}\{S_t = k\}/n_k$, $v_k = n_k^{-1/2}$. However, ρ_k is a separate within-state selection quantity: low confusion mass alone does not bound it. If $\beta_k = \beta_0 + \delta_k$, then $\Delta_{k\ell} = \|\delta_\ell - \delta_k\| \leq \|\delta_\ell\| + \|\delta_k\|$. For s-sparse departures bounded coordinatewise by B, the deterministic contamination term is at most $2R^2 \eta_k \sqrt{s} B$. If there are also state-specific stable coefficients, their difference must be added rather than hidden in η_k.

Sufficient consistency conditions under a bounded Gram inverse are $\eta_k \max_{\ell \neq k} \Delta_{k\ell} \to 0$, $v_k \sqrt{p \log(Kp)} \to 0$, and $\rho_k \sqrt{N} \to 0$; none follows from ARI. For $x_t = 1$ and one true state, select weights uniformly over samples with $\epsilon_t > 0$, when nonempty. Misassignment is zero, yet the selected noise mean converges to $\sigma\sqrt{2/\pi}$. The selection term retains this target-dependence risk.

Exact relationship to the implemented model. For fixed state k and target j, the native decoder has input $z_{k,j} = p \odot A_{k,:,j}$ and mean $\alpha_{kj} f_j(z_{k,j}) + \beta_{kj}$. At a fixed operating point its parent derivative is

$$J_{k,j,i} = \alpha_{kj}\, \partial_i f_j(z_{k,j})\, A_{\text{global},ij} G_{k,ij}.$$

A mask gives a zero derivative; a nonzero gate guarantees neither a nonzero derivative nor a shared sign. Derivatives vary with the operating point. Applying Eq. (18) to a local Taylor approximation requires a decoder-specific Hessian bound; the tanh decoder is not globally the fixed linear model.

Neural learning couples the posterior, logits, gates, decoder and affine terms through per-state masks, straight-through gradients, spectral normalization and regularization. It neither enumerates a common hard support nor supplies a noise-independent split. The theorems therefore describe explicit estimators, not the trained neural optimizer. Total regime exposure, segment length, overlapping windows and independent observations remain distinct. The diagnostic variant in Appendix C removes VIS spectral constraints and changes input scaling; bounds requiring those constraints or constants are not transferred to it.

A.6 Within-State Response Consistency and Mechanism Mixing

Within one mechanism, responses can still vary with the input. Define $R_f(x,h) = f(x+h) - f(x)$ and $q(v) = \|v\|_1/d$. The all-output pair loss is $\|N_\epsilon(R_1) - N_\epsilon(R_2)\|_2^2/d$, where $N_\epsilon(v) = v/\max\{q(v),\epsilon\}$.

Proposition A.4 (Quantitative within-state consistency)**.** *Suppose J_f is L_J-Lipschitz on a convex domain containing x, y and their common perturbation segments. If $\|h\| \le \delta$ and $\|x-y\| \le r$, then*

$$\|R_f(x,h) - R_f(y,h)\|_2 \le L_J \delta r.$$

Suppose both learned responses approximate the corresponding true responses within e_R, have norm at most M_R, and satisfy $q(\widehat{R}) \ge c_{\min} > \epsilon$. Their CNC pair loss is at most

$$\frac{1}{d}\left(\frac{1}{c_{\min}} + \frac{M_R}{\sqrt{d}\,c_{\min}^2}\right)^2 (L_J\delta r + 2e_R)^2.$$

Proof. The fundamental theorem of calculus gives

$$R_f(x,h) - R_f(y,h) = \int_0^1 [J_f(x+th) - J_f(y+th)]h\,dt.$$

The Jacobian condition bounds the integrand norm by $L_J\|x-y\|\|h\|$, proving the raw-response bound. For two learned responses v, w, write

$$\frac{v}{q(v)} - \frac{w}{q(w)} = \frac{v-w}{q(v)} + w\left[\frac{1}{q(v)} - \frac{1}{q(w)}\right].$$

Since $|q(v) - q(w)| \le \|v-w\|_2/\sqrt{d}$, the normalized difference is at most $(c_{\min}^{-1} + M_R/(\sqrt{d}\,c_{\min}^2))\|v-w\|_2$. The triangle inequality adds the two approximation errors. Squaring and dividing by d gives the pair-loss bound. □

The bound explicitly separates mechanism curvature, input separation and response strength. For the unchanged mechanism $f(u) = (u^2, u)^\top$, the response to fixed $h \ne 0$ is $(2uh + h^2, h)^\top$; its component ratio changes with u. Robust data-scale normalization alone therefore cannot ensure within-state response agreement.

A.6.1 Mechanism-Mixing Lower Bound

Lemma A.1 (Conditional mixed-pair penalty)**.** *Suppose a discriminative intervention is selected with probability at least $p_h > 0$ and, conditional on that event, a cross-mechanism pair is drawn with probability at least $p_m > 0$. Suppose the true normalized responses for every pair in this joint event are separated by at least γ, and each learned normalized response is within $e < \gamma/2$ of its corresponding true response. Then the expected native all-output pair loss is at least $p_h p_m(\gamma - 2e)^2/d > 0$.*

Proof. The triangle inequality lower-bounds the learned response distance on the joint event by $\gamma - 2e$. Multiply its squared value divided by d by the joint-event probability, which is at least $p_h p_m$. Contributions outside the event are nonnegative. The bound applies to the specified pair distribution; averaging over other states additionally requires their sampling mass. Positive marginal event probabilities without a conditional bound do not suffice. □

The lemma does not prevent oversegmentation: fragmentation can remove eligible mixed pairs. Singleton-state fallback must be included in the actual pair distribution.

A.6.2 Boundary Cases: Empty Graph, Bypass, and Near-Silent Response

Zero responses have zero floored CNC loss. Positive gain rescaling cancels only above the floor, and proportional response vectors need not be distinguishable. Neither a support budget, sparsity penalty nor detached NEC supervision excludes all silent-response, empty-graph or bypass solutions.

A.6.3 Normalization values and radial updates

Forward normalization and its derivative play different roles in optimization, as demonstrated for LayerNorm by Xu et al. (2019). Stop-gradient procedures must also be distinguished from gradient descent on their displayed forward objective (Ponce et al., 2026). Here the object is specifically the mean-absolute normalization of a signed intervention response.

Write $c(r) = \|r\|_1/d$, $N_\epsilon(r) = r/\max\{c(r), \epsilon\}$, and $\ell(r, s) = \|N_\epsilon(r) - N_\epsilon(s)\|_2^2/d$. For this derivative comparison, SG denotes CNC's stop-gradient rule: it differentiates the numerators with the numerical denominators fixed. FG denotes the experimental alternative that differentiates both. Their forward values agree.

Proposition A.5 (Radial derivative with fixed pairs)**.** *Fix the pair, working points, perturbation, state weights and discrete choices. Suppose $c(ar), c(as) > \epsilon$ on an open interval of positive scales a. Then $\ell(ar, as)$ is constant on that interval, whereas its SG backward value and FG derivative satisfy*

$$G_{\mathrm{SG}}(a) = 2\ell(ar, as)/a, \qquad G_{\mathrm{FG}}(a) = 0.$$

At the current responses,

$$r^\top \nabla_r^{\mathrm{SG}} \ell + s^\top \nabla_s^{\mathrm{SG}} \ell = 2\ell, \qquad r^\top \nabla_r^{\mathrm{FG}} \ell = s^\top \nabla_s^{\mathrm{FG}} \ell = 0.$$

Proof. Positive homogeneity gives $c(ar) = ac(r)$, hence $N_\epsilon(ar) = r/c(r)$ above the floor. Put $u = r/c(r)$ and $v = s/c(s)$. Holding denominators fixed in the backward calculation gives $\nabla_r^{\mathrm{SG}} \ell = 2(u-v)/(dc(r))$ and $\nabla_s^{\mathrm{SG}} \ell = -2(u-v)/(dc(s))$. Their radial inner products sum to $2\|u-v\|^2/d = 2\ell$; at scale a, the chain rule for the numerators contributes $2\ell/a$. For full differentiation, the Jacobian at nonzero coordinates is

$$J_N(r) = \frac{I}{c(r)} - \frac{r\ \mathrm{sign}(r)^\top}{d\, c(r)^2}.$$

Since $\mathrm{sign}(r)^\top r = dc(r)$, $J_N(r)r = 0$, proving each FG radial identity. This L1-based Jacobian is generally nonsymmetric; it is not the orthogonal projection associated with L2 normalization. At a zero coordinate the implementation uses the absolute-value subgradient zero. The same radial identity holds because that coordinate contributes zero to the inner product. □

Below the floor, $J_N(r) = I/\epsilon$. Away from floor equality, write $u = N_\epsilon(ar)$ and $v = N_\epsilon(as)$. The common-scale FG derivative is

$$\frac{2}{da}\left[\mathbf{1}_{c(ar)<\epsilon} u^\top (u - v) + \mathbf{1}_{c(as)<\epsilon} v^\top (v - u)\right].$$

Thus both floored endpoints give $2\ell/a$, and a mixed pair need not have zero radial derivative. At equality the function may have different one-sided derivatives; the implemented clamp passes the derivative of its input at equality. These cases, including the zero vector, are recorded separately in the numerical checks. Fixed nonnegative pair weights preserve the identities by averaging; weights that depend on a require additional terms.

Proposition A.6 (Conditional nonstationarity of a correct mechanism)**.** *Suppose the representation contains $f_a = af_*$, $a > 0$, with $Y = f_*(X) + \eta$ and $\mathbb{E}[\eta \mid X] = 0$. Assume finite second moments, $Q = \mathbb{E}\|f_*(X)\|^2/d > 0$, a fixed state/pair distribution, and responses above the floor throughout a neighborhood of $a = 1$. Let $C_0 = \mathbb{E}\ell(R_{f_*}(X, h), R_{f_*}(X', h)) > 0$, with differentiation and expectation interchangeable. If other regularizers have zero radial derivative, then*

$$G_{\mathrm{SG,total}}(a) = 2Q(a - 1) + \frac{2\lambda_{\mathrm{CNC}} C_0}{a}.$$

Consequently $a = 1$ is not stationary for this SG update field when $\lambda_{\mathrm{CNC}} > 0$. FG contributes zero CNC derivative in this direction.

Proof. Expanding $\mathbb{E}\|Y - af_*(X)\|^2/d$ and conditioning on X removes the cross term, leaving $Q(a-1)^2 + \mathbb{E}\|\eta\|^2/d$. Its derivative is $2Q(a-1)$. Apply Proposition A.5 and average over the fixed distribution. At $a = 1$ the SG term is $2\lambda_{\mathrm{CNC}} C_0 > 0$. □

In this scalar model, stationary scales solve $a^2 - a + \lambda_{\mathrm{CNC}} C_0/Q = 0$, provided the discriminant is nonnegative and the positive roots remain in the nonfloor interval. These stationary points describe the stated scale family. The full model also updates state probabilities and pairs, with spectral constraints, clipping and Adam preconditioning.

Working-point mismatch is a separate issue. For the additive sine mechanism and a source-i perturbation h_i, the physical response is $2\sin(h_i)\cos(2x_i + h_i)W[k, i, :]$. Its sign can reverse within one mechanism, making $C_0 > 0$. FG removes the radial semigradient term above the floor but leaves this forward pair discrepancy unchanged. The consistency bound in Proposition A.4 still uses its input-distance, curvature and response-strength assumptions.

A.7 Factorization Freedom and Canonical Representatives

For an effective edge $e_k = am_k g_k$, $a, g_k \in (0, 1)$ and $m_k \in \{0, 1\}$, reciprocal rescaling can leave every e_k unchanged while changing the shared factor. A fixed export rule does not remove this freedom.

Proposition A.7 (Canonical effective-product representation)**.** *For any nonnegative effective products $e_1, \ldots, e_K$ with $u = \max_k e_k > 0$, the representation $a^* = u$, $h_k = e_k/u$ is the unique representation satisfying $e_k = a^* h_k$, $h_k \geq 0$, and $\max_k h_k = 1$. For $u = 0$, choose $a^* = h_k = 0$ by convention.*

Proof. The displayed construction preserves each product and has maximum normalized gate one. Conversely, taking maxima of $e_k = a^* h_k$ under the constraint gives $u = a^*$, after which every $h_k = e_k/u$ is determined. The zero convention specifies the otherwise undetermined gates when all products vanish. □

This construction defines a canonical *output*, leaving training unchanged. It should be applied to effective products after support selection: normalizing gates before top-k can change their ordering. For example, gate values 0.4 and 0.3 with cross-state maxima 0.9 and 0.3 become 0.444 and 1.

B PUBLIC BENCHMARK SUPPLEMENT

The public benchmarks cover dynamic graph recovery in DREAM3 and fMRI, and static graph recovery in Sachs and SynTReN. Arrows indicate the preferred metric direction; bold identifies the best displayed mean, not statistical significance.

B.1 DREAM3 PROTOCOLS AND AUXILIARY COMPARISON

Table 6: DREAM3 Size100 structural ranking over all 9,900 directed nonself pairs. Values are percentages.

AUROC ↑ Method	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3	Macro
PCDM-KEEP	54.52±0.50	52.65±1.15	51.96±0.27	51.04±1.07	50.55±0.46	52.14±0.42
PCDM-RANDOM†	51.65±2.06	50.77±2.11	50.40±1.72	51.82±0.56	50.69±0.12	51.07±0.59
PCDM-INFO	54.72±1.69	53.19±1.36	56.51±1.42	53.67±0.60	51.48±0.29	53.92±0.18
NTS-NOTEARS	**58.63±0.57**	57.31±0.23	**59.59±0.25**	**54.41±0.06**	**52.29±0.08**	**56.44±0.16**
DYNOTEARS	58.23	**58.41**	58.15	52.90	51.98	55.93
PCMCI-ParCorr	52.30	48.56	57.72	51.38	50.35	52.06
VARLiNGAM	57.17	53.99	54.82	51.79	51.28	53.81

AP ↑ Method	Ecoli1	Ecoli2	Yeast1	Yeast2	Yeast3	Macro
PCDM-KEEP	1.82±0.36	2.02±0.35	2.58±0.55	4.13±0.15	5.87±0.20	3.28±0.11
PCDM-RANDOM†	1.73±0.22	1.58±0.36	2.24±0.37	4.39±0.19	5.87±0.11	3.16±0.15
PCDM-INFO	1.86±0.23	2.27±0.25	3.58±0.63	4.84±0.24	6.20±0.16	3.75±0.22
NTS-NOTEARS	4.67±0.06	5.30±0.08	6.47±0.04	6.30±0.05	6.67±0.02	5.88±0.03
DYNOTEARS	**7.71**	**6.97**	**8.03**	**7.17**	**7.06**	**7.39**
PCMCI-ParCorr	2.18	1.27	3.02	4.56	5.93	3.39
VARLiNGAM	2.49	1.75	3.14	4.99	5.91	3.65

PCDM and NTS-NOTEARS entries are mean ± sample SD over five training seeds. Macro first averages the five fixed networks within each seed, then summarizes those five averages. DYNOTEARS, PCMCI-ParCorr and VARLiNGAM have one deterministic output per network and no seed SD; the five requested repeats for each of the latter two produce identical score arrays. †RANDOM is a post hoc exploratory control. AP denotes average precision. Bold marks the largest mean for each network and metric.

External structural ranking. Table 6 compares the three PCDM restart conditions with NTS-NOTEARS, DYNOTEARS, PCMCI-ParCorr and VARLiNGAM. INFO has the largest mean AUROC and AP among the three PCDM conditions on each network. NTS-NOTEARS leads macro AUROC, while DYNOTEARS leads macro AP; both exceed INFO on both macro metrics. PCMCI-ParCorr and VARLiNGAM each return identical score arrays across five repeated fits per network, so their entries represent one deterministic result per network without a seed SD.

INFO's paired gains over KEEP are 1.77 AUROC points (95% CI $[1.22, 2.32]$) and 0.47 AP points ($[0.26, 0.67]$), both passing the two-endpoint Holm correction. Relative to RANDOM, its gains are 2.85 and 0.59 points, with exploratory intervals $[2.32, 3.38]$ and $[0.37, 0.81]$. INFO's selected validation reconstruction MSE is higher than KEEP's in 23 of 25 pairs. This reconstruction uses offline posterior context rather than strictly causal one-step prediction.

Each Size100 network contains 46 independent 21-point trajectories over 100 genes. A fixed shuffle (seed 20260916) assigns 23 trajectories to training, 11 to validation and 12 to test. Means and population standard deviations use training observations only. Each length-32 window contains one right-padded trajectory with 20 valid lag-one targets; no window crosses a trajectory boundary. The residual-guided learner uses three states, batch size 16, hard Top-5 and Adam learning rate 0.001. Shared-affine weight is 0.001; sparsity weight is 0.005 for Ecoli2 and Yeast3, and 0.001 for the other networks. DAG and NEC terms are zero. Each 90-epoch stage has 180 decoder and 170 posterior updates; the complete two-stage learner has 360 and 340, respectively.

Matched learning and selection. Each network–seed pair shares a validation-selected 90-epoch base model. KEEP preserves the incoming gate assignment; INFO transports the same incoming gate-logit multiset to the ranking from a training-fitted additive spline in the base model's state order. The screening fit considers all 99 nonself inputs per target and selects among ridge coefficients $\{0.001, 0.1, 10\}$ by observation-validation reconstruction, without refitting. The final neural graph retains hard Top-5. KEEP, RANDOM and INFO reset the Adam states and receive 90 further epochs. The first five epochs adapt decoder parameters with the base posterior and frozen structural and posterior parameters; the remaining 85 resume joint learning. Checkpoint selection uses the earliest exact minimum observation-validation reconstruction, including epoch 0. The auxiliary comparison adds the signed CNC-SG term with coefficient 0.001 and posterior gradients connected.

Randomized support control. RANDOM preserves each state–target column's complete nonself gate-logit multiset and matches INFO's overlap with the base Top-5 support. Retained and replacement parents are sampled without replacement. The selected Top-5 and remaining candidates are shuffled separately, then assigned the descending logit values. Network indices $n = 1, \ldots, 5$ follow Ecoli1, Ecoli2, Yeast1, Yeast2 and Yeast3; s denotes the training seed. The private support seed is the first uint32 output of `SeedSequence([820701,n,s]).generate_state(1)`, passed as an integer to NumPy's `default_rng`. The shared training seed uses the same uint32 conversion with domain 820700. These 25 restarts use the same bases, windows, scaler, training seeds and selection rule, with CNC, NEC and DAG weights zero. RANDOM is a post hoc control conditional on the observed KEEP/INFO outcomes; its paired intervals are exploratory and unadjusted.

Table 7: Supplementary randomized-control macro contrasts on DREAM3. Differences and ordinary paired 95% t_4 intervals are in percentage points.

Contrast	Metric	Difference [95% CI]
INFO − RANDOM	AUROC ↑	$+2.85$ $[+2.32, +3.38]$
INFO − RANDOM	AP ↑	$+0.59$ $[+0.37, +0.81]$
RANDOM − KEEP	AUROC ↑	-1.08 $[-2.09, -0.06]$
RANDOM − KEEP	AP ↑	-0.12 $[-0.36, +0.12]$

Each contrast uses five paired seed-level averages over the five fixed networks. Positive differences favor the first method. These supplementary contrasts are separate from the prespecified Holm families in Table 10.

External comparison and inference. DYNOTEARS fits an instantaneous DAG W and an unconstrained lag matrix A using trajectory-safe lag pairs and deterministic initialization. Six candidates per network combine $\lambda_W \in \{0.001, 0.01, 0.1\}$ with $\lambda_A \in \{0.01, 0.1\}$. Selection minimizes observation-validation generative prediction error, evaluated through a linear solve; no selected model is refitted. The selected coefficients are $\lambda_W = \lambda_A = 0.1$ for all five networks. All five selected solutions have successful terminal inner solves and acyclicity residual below 10^{-8}; one unselected candidate has an unsuccessful terminal inner solve. Structural ranking uses all 9,900 directed nonself pairs: the maximum state-specific effective weight for PCDM and the prespecified $|W + A|$ score for DYNOTEARS. The five PCDM training seeds are 1001, 1003, 1007, 1009 and 1013. Macro inference conditions on the five fixed networks and uses five seed-level paired averages. Ordinary t_4 confidence intervals and Holm-adjusted tests are reported in Table 10.

Additional external methods. PCMCI-ParCorr (Runge et al., 2019) uses Tigramite 5.2.10.1 and its native multiple-trajectory interface at lag one. PC1 selects among thresholds $\{0.01, 0.05, 0.1, 0.2\}$ separately for each target by training AIC, followed by full MCI testing. We use analytic partial-correlation tests, native `max_combinations=1`, unrestricted conditioning-set dimensions and no FDR adjustment. The prespecified continuous score is the negative unadjusted lag-one MCI p-value, rather than a thresholded graph or effect magnitude. The native two-lag cutoff yields 437 MCI samples (23×19, target times 2–20); PCDM and the lag-one prediction methods use 460 training targets (23×20, times 1–20).

VARLiNGAM (Hyvärinen et al., 2010) uses lag one, native `pwling` DirectLiNGAM and adaptive-Lasso pruning, with a single fixed configuration and no hyperparameter grid. Validation prediction is descriptive and does not change this configuration. To accommodate independent trajectories, lagged rows are formed within each trajectory before pooling the no-intercept regression, residual

and pruning designs. Native causal ordering and pruning are retained with this trajectory-boundary adaptation. The prespecified score is $|B_0 + B_1|^\mathsf{T}$, converting native target–source matrices to source–target orientation while retaining both instantaneous and lagged terms.

NTS-NOTEARS (Sun et al., 2023) uses the authors' nonlinear implementation (commit `e1ba34b784f9eec3d2139904589d2fd1ed4ea7ad`), lag one, dimensions $[100, 10, 1]$ and float64. Its native SciPy L-BFGS-B solver and augmented-Lagrangian procedure enforce a contemporaneous DAG, with $\lambda_2 = 0.01$, $h_{\text{tol}} = 10^{-8}$, $\rho_{\max} = 10^{16}$ and at most 100 outer iterations. Reconstruction losses are accumulated over separate trajectories; regularization and the DAG term are each added once. With development seed 9001, $\lambda_1 \in \{0.0005, 0.0015, 0.005, 0.015, 0.05, 0.15\}$ is selected per network by validation conditional-reconstruction MSE, breaking ties by the earlier candidate index. All five networks select 0.015 and are refitted from scratch with seeds 1001, 1003, 1007, 1009 and 1013. The exported score is the elementwise $\max(W_0, W_1)$ of the native nonnegative instantaneous and lagged group strengths, in source–target orientation. Its conditional reconstruction uses current and lagged observations, rather than only past observations. Methods share the observations and edge candidates while retaining their native solvers, selection rules and information budgets.

All methods rank the 9,900 directed nonself pairs and report AUROC and average precision. PCDM uses the maximum state-specific effective weight. PCMCI-ParCorr and VARLiNGAM each have five requested executions per network, with identical score arrays within each network. They contribute one deterministic network result and no seed SD to Table 6.

External computation cost. The 25 PCMCI-ParCorr executions have a cumulative recorded wall time of 33.43 minutes, including 4.24 minutes of additional computation; only wall time is reported. The 25 VARLiNGAM executions use 22.21 minutes of cumulative wall time and 21.77 minutes of CPU time. DYNOTEARS uses 30 candidates, totaling 57.00 minutes of wall time and 56.38 minutes of CPU time, with no selected-model refit. These totals include actual repeated computations and describe process intervals on the workstation. The deterministic repeats do not add independent statistical replicates.

Table 8: NTS-NOTEARS computation across all fits. Times sum all measured process intervals, including additional computation, and exclude queue waiting.

Method	Development (30 fits)	Formal (25 fits)	Total (55 fits)
NTS-NOTEARS	26.71	17.13	43.84

Table 9: DREAM3 macro results with and without the auxiliary CNC-SG term. Scores are percentages.

Condition	AUROC ↑	AP ↑
KEEP	52.14 ± 0.42	3.28 ± 0.11
INFO	$\mathbf{53.92 \pm 0.18}$	$\mathbf{3.75 \pm 0.22}$
KEEP+CNC	52.16 ± 0.42	3.30 ± 0.09
INFO+CNC	53.87 ± 0.17	3.74 ± 0.20

Mean ± sample SD uses the five seed-level macro averages defined in Table 6. Each condition receives the same second 90-epoch stage. Bold marks the largest mean.

Computational cost. The complete INFO procedure includes base training, screening and neural restart. For the controlled additive and interaction families, recorded neural intervals total 124.18 ± 6.21 and 92.79 ± 1.92 minutes; screening takes 0.85 ± 0.06 and 0.82 ± 0.01 CPU seconds. On DREAM3, the second stage averages 74.26 minutes for KEEP and 77.89 minutes for INFO. These DREAM3 times cover the restart stage and exclude base-model training. Timings come from a shared workstation. Dense masked-input operations dominate the decoder; a small parent budget does not make its kernels sparse.

Table 10: Prespecified DREAM3 paired contrasts. Differences and intervals are in percentage points.

Contrast	Metric	Difference [95% CI]	Raw p	Family p	All-ten p
$I - K$, CNC off	AUROC	$+1.772$ $[+1.220, +2.325]$	0.00088	0.00175	0.00803
$I - K$, CNC on	AUROC	$+1.714$ $[+1.192, +2.236]$	0.00080	0.00642	0.00803
CNC on−off, KEEP	AUROC	$+0.012$ $[-0.006, +0.029]$	0.14407	0.86441	0.86441
CNC on−off, INFO	AUROC	-0.047 $[-0.136, +0.043]$	0.22348	0.89390	0.89390
Interaction	AUROC	-0.058 $[-0.150, +0.034]$	0.15319	0.86441	0.86441
$I - K$, CNC off	AP	$+0.468$ $[+0.261, +0.674]$	0.00325	0.00325	0.02602
$I - K$, CNC on	AP	$+0.432$ $[+0.220, +0.644]$	0.00479	0.03356	0.03356
CNC on−off, KEEP	AP	$+0.021$ $[-0.037, +0.078]$	0.38147	0.89390	0.89390
CNC on−off, INFO	AP	-0.015 $[-0.112, +0.083]$	0.69681	0.89390	0.89390
Interaction	AP	-0.035 $[-0.115, +0.045]$	0.28923	0.89390	0.89390

I denotes INFO and K denotes KEEP. Each contrast uses five paired seed-level macro differences. Intervals are ordinary two-sided t_4 intervals, not simultaneous intervals. Family p adjusts the two primary CNC-off $I - K$ endpoints together and the remaining eight tests together; All-ten p adjusts all ten by Holm. Interaction is $(I - K)_{\text{CNC on}} - (I - K)_{\text{CNC off}}$.

Table 11: DREAM3 second-stage cost and validation-selected learning epoch.

Condition	Wall (min)	CPU (min)	Peak (MiB)	Epoch: median [range]	Epoch 0
KEEP	74.26 ± 10.75	27.45	161.2	3 [0, 56]	5/25
RANDOM	46.40 ± 31.50	22.18	161.2	67 [40, 87]	0/25
INFO	77.89 ± 13.07	27.71	161.2	69 [41, 90]	0/25
KEEP+CNC	81.49 ± 13.25	29.90	247.3	3 [0, 75]	5/25
INFO+CNC	85.46 ± 15.28	30.65	247.3	73 [36, 90]	0/25

Wall is mean $\pm$ sample SD of 25 training intervals; CPU and peak allocated PyTorch memory are means. Conditions were timed during different periods on a shared workstation; these measurements do not establish relative algorithmic speed. All runs complete 90 epochs; observation-validation selection includes epoch 0. Each full procedure uses a 90-epoch base plus this 90-epoch stage (360 mechanism-objective and 340 posterior updates). Reported wall times cover the restart stage only. INFO adds three screening candidates per network–seed pair: wall 1.35 ± 0.10 s and CPU 1.33 ± 0.10 s. DYNOTEARS uses six candidates per network (11.40 min mean total candidate wall), selects by observation-validation MSE, and performs no refit.

B.2 fMRI, Sachs and SynTReN

fMRI simulation. Table 12 reports the network simulation of Smith et al. (2011). With NEC off, CNC weight 10^{-3} slightly improves AUROC, F1 and SHD over CNC-off, but lowers AUPRC. The configurations use indegree four, sparsity/DAG weights 0.0075/0.1, perturbation fraction 0.008, eight pairs, self-output CNC, window 64 and batch 96, with NEC disabled.

Table 12: fMRI simulation results.

CNC weight	AUROC ↑	AUPRC ↑	F1 ↑	SHD ↓
0	0.7613 ± 0.0262	0.4121 ± 0.0558	0.3333 ± 0.0809	12.10 ± 1.79
5×10^{-4}	0.7639 ± 0.0277	$\mathbf{0.4149 \pm 0.0533}$	0.3238 ± 0.0666	12.30 ± 1.49
10^{-3}	$\mathbf{0.7643 \pm 0.0275}$	0.4093 ± 0.0527	$\mathbf{0.3429 \pm 0.0803}$	$\mathbf{11.80 \pm 1.87}$

Sachs static transfer. Table 13 compares PCDM with published methods on the protein-signaling benchmark (Sachs et al., 2005). PCDM has higher SHD and lower F1 than the strongest listed methods. Its settings use indegree 0, STE enabled, sparsity/DAG weights 0.003/0.05, CNC weight 2×10^{-4}, perturbation 0.0015, four pairs and self-output comparison.

Table 13: Sachs results. PCDM reports mean ± sample SD over seeds 2, 5, 7, 8, 9 and 10; other rows are from Xu et al. (2024), Table 1 on page 8.

Method	SHD ↓	F1 ↑	Method	SHD ↓	F1 ↑
NOTEARS	12.0 ± 0.00	0.387 ± 0.000	SCORE	12.0 ± 0.00	0.444 ± 0.000
GOLEM	17.0 ± 0.00	0.421 ± 0.000	DiffAN	12.2 ± 0.98	0.434 ± 0.078
GraNDAG	13.2 ± 0.75	0.373 ± 0.064	DAGuerreotype	17.9 ± 0.54	0.118 ± 0.034
CAM	12.0 ± 0.00	0.444 ± 0.000	CaPS	$\mathbf{11.0 \pm 0.00}$	$\mathbf{0.500 \pm 0.000}$
VI-DP-DAG	42.6 ± 1.36	0.340 ± 0.037	PCDM	17.0 ± 0.63	0.3333 ± 0.0263

SynTReN source-network stress tests. We use static linear-SEM observations defined on the Ecoli and yeast source networks distributed with SynTReN (Van den Bulcke et al., 2006) (Table 14). Joint capacity/sparsity changes increase SHD and graph density, alongside higher F1; disabling STE further worsens recovery.

Table 14: Stress tests on SynTReN source networks. Edge counts are predicted/true.

Dataset	Setting	F1 ↑	SHD ↓	Edge counts
Ecoli	Fixed protocol	0.176	**714.5**	401.2/519
Ecoli	Oracle threshold	0.161	812.4	501.5/519
Ecoli	Capacity/sparsity alternative	**0.250**	867.7	740.7/519
Yeast	Fixed protocol	0.116	**1458.7**	652.0/1082
Yeast	Capacity/sparsity alternative	**0.121**	1893.3	1211.7/1082
Yeast	STE off	0.019	2295.0	1330.0/1082

Ecoli fixed/oracle rows average seeds 1–10; the other rows average seeds 1, 2 and 5. Oracle thresholding uses ground truth and is diagnostic. Fixed-protocol rows use $k = 1$, sparsity/DAG weights 0.08/0.08; joint configuration alternatives use $k = 2$, 3.00/0.20, with STE disabled only in the last row.

C Controlled Data and State-Specific Recovery

The controlled experiments distinguish an edge's presence in the union graph from its assignment to the mechanism that uses it. The same graph/data/training tuples support the known-state, residual-guided and structured-state comparisons.

C.1 A CONTROLLED ADDITIVE MECHANISM AND INTERVENTION-RESPONSE GATE

For each of 20 target variables, we sample five distinct non-self parents: two CORE parents active in all three states, and three PRIVATE parents assigned one-to-one to the three states. Every state therefore has in-degree three at every target. Each graph contains 40 CORE edges, 60 PRIVATE edges, and 280 ABSENT edges among the 380 directed non-self pairs. The sine-additive generating mechanism is

$$\begin{aligned} x_{t,j} &= f_{s_t,j}(x_{t-1}) + \epsilon_{t,j}, \quad \epsilon_{t,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.1^2), \\ f_{k,j}(x) &= \sum_i W_{k,ij} \sin(2x_i). \end{aligned} \tag{22}$$

Active coefficients have random signs and magnitudes sampled uniformly from $[0.58, 0.62]$; a CORE coefficient is shared across states. The states recur in a fixed graph-specific cyclic order, with 64 targets per segment. After 384 burn-in steps from a zero initial condition, we retain 6,144 observations.

Signal strength is checked at the level of the complete generating function. On a separate calibration trajectory, set $\delta_i = 0.1(Q_{0.95}(x_i) - Q_{0.05}(x_i))$, using previous-source values, and define

$$\eta_{k,ij}(x) = \frac{|f_{k,j}(x + \delta_i e_i) - f_{k,j}(x)| + |f_{k,j}(x - \delta_i e_i) - f_{k,j}(x)|}{2(0.1)}. \tag{23}$$

Every active-coordinate median must be at least 1.5 noise SD; the CORE and PRIVATE class medians must lie in $[1.5, 2.5]$, with their ratio in $[0.75, 1.3334]$. Calibration and an independent innovation audit use the same perturbation values and exclude the first target after each switch. All five fixed graph/weight draws pass, without resampling: the smallest audited active-coordinate median is 1.6338 and class medians span 2.0077–2.0685. Inactive private responses are zero. A third innovation stream provides formal observations, whose evaluation includes all valid targets.

C.2 EVALUATION OBJECTS AND KNOWN-STATE CONTROLS

Let $\mathcal{E}_P$ denote the 60 PRIVATE edges, z_e their true owners, and $a_{k,e}$ the scores from the actual decoder effective graph $A^{(k)}_{\text{eff}}$. PRIVATE×state AP pools the 180 scores against $y_{k,e} = \mathbf{1}\{k = z_e\}$: 60 positives and 120 negatives corresponding to the same edges in the wrong states. Equal scores are grouped when computing $\text{AP} = \sum_b (R_b - R_{b-1}) P_b$, where P_b and R_b are precision and recall after adding the bth distinct score group in descending order. For $H_e = \arg\max_k a_{k,e}$, the other two primary summaries are

$$\text{Owner} = \frac{1}{60} \sum_{e \in \mathcal{E}_P} \frac{\mathbf{1}\{z_e \in H_e\}}{|H_e|}, \qquad \overline{m} = \frac{1}{60} \sum_{e \in \mathcal{E}_P} \left(a_{z_e,e} - \max_{k \neq z_e} a_{k,e} \right). \tag{24}$$

All 60 edges enter both denominators. An all-zero edge receives owner credit $1/3$ and margin zero, not a correct unique assignment. Union AP uses $\max_k a_{k,e}$ over all 380 non-self candidate pairs, with 100 union-positive edges. It answers whether an edge exists anywhere and cannot establish correct state ownership.

We use five fixed (graph, data, training-stream) tuples: $(11, 101, 1001)$, $(23, 103, 1003)$, $(37, 107, 1007)$, $(53, 109, 1009)$, and $(71, 113, 1013)$. Chronological train/validation/test fractions are $1/2, 1/4, 1/4$; standardization uses training observations only. The direct-reconstruction comparison has five runs per structure with given states and five per three-state structure with inferred states; the state-agnostic Single-graph runs are reused. In inferred-state evaluation, one global permutation is fixed using validation soft-confusion agreement. All runs use 90 epochs, a per-state parent budget of three, and CNC and NEC disabled; checkpoint selection uses validation reconstruction MSE within the budget. Oracle runs receive the target-state label directly.

C.3 RESIDUAL-GUIDED LEARNING

Responsibilities, updates and readout. The residual teacher uses Gaussian log emissions

$$e_{t,k} = -\frac{1}{2} \left[\frac{\|x_t - \mu_t^{(k)}\|_2^2}{v\tau_e} + d \log(2\pi v \tau_e) \right], \tag{25}$$

with one estimated variance v and temperature τ_e shared across states. Its state marginals are detached for Eq. (6). The sparsity penalty is the elementwise absolute mean of the state-mean effective graph after zeroing the diagonal, with the full d^2 denominator.

Each window starts with a uniform state prior. The initial transition matrix has $P_{kk} = 0.9$ and $P_{kj} = 0.1/(K-1)$ for $j \neq k$. Forward–backward inference uses valid tokens only, with no transitions across windows, trajectories or splits. At each epoch, teacher marginals and pair responsibilities are computed once and detached. After the neural updates, the transition rows are normalized from cached training pair counts plus pseudocounts 1 on the diagonal and 0.1 elsewhere. The updated decoder and frozen responsibilities estimate $v = \mathrm{mean}_t \sum_k r_{t,k} \|x_t - \mu_t^{(k)}\|^2/d$, clipped to $[10^{-4}, 10]$; its initial value is one. The temperature is two at epoch 6, decreases linearly to one at epoch 15, and remains one.

Bootstrap responsibilities are obtained by collecting initial Transformer logits over valid training exposures in evaluation mode, centering each state column by its exposure mean, and applying softmax. This bank is held fixed for epochs 1–5, which update only the decoder; posterior updates start at epoch 6. Both Adam optimizers use $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay and separate norm clipping at one. Training uses float32. The neural posterior selects checkpoints by validation reconstruction; teacher agreement is diagnostic. Responsibility-weighted branch MSE includes the within-responsibility variance of branch means in addition to their weighted-mixture MSE, so it is a different objective from direct mixture reconstruction.

For the controlled data, the posterior is a one-layer Transformer with width 32, four heads, feed-forward width 64 and no dropout. VIS has two width-32 tanh layers with spectral normalization and unit input gain. Windows have length 32 and train/evaluation strides 16/32, remain within their chronological splits, and use batch size 16. Both Adam optimizers use learning rate 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay and separate norm clipping at one. Shared-affine and effective-edge sparsity weights are both 0.001. CNC, NEC and DAG terms are zero except in explicitly named auxiliary comparisons. Each 90-epoch run has 1,080 decoder and 1,020 posterior updates; neural-q validation MSE selects the earliest minimum checkpoint.

Given states, the PCDM-minus-Independent intervals are $[-0.0198, 0.0519]$ for PRIVATE AP, $[-0.0074, 0.0585]$ for owner and $[-0.0144, 0.0692]$ for margin. With direct mixture inference, PCDM's state ARI is 0.00098. Residual-guided learning improves state recovery, but its corresponding localization intervals, $[-0.0939, 0.0994]$, $[-0.1092, 0.1003]$ and $[-0.0530, 0.0758]$, also include zero.

C.4 The structured-state diagnostic variant

Unique targets and temporal inference. Each physical target is counted once. When 32-step windows overlap, raw state logits and state-specific predictive means are averaged equally over their exposures before softmax or residual computation. Continuous components are defined by trajectory, split, and consecutive target indices; window seams do not split them. Training uses stride 16, with 3,071 unique targets; validation and test each contain 1,535. The duration distribution is

$$D = 1 + \mathrm{NB}(2, p), \quad \Pr(D = d) = dp^2(1-p)^{d-1}\ (d \geq 1), \quad \mathbb{E}[D] = 2/p - 1. \tag{26}$$

The shared initial $p = 2/11$ implies mean duration 10, chosen independently of the generating segment length. Initial semantic-state probabilities are uniform. The between-segment transition matrix has zero diagonal and initially equal off-diagonal entries. Training updates it using expected transitions with off-diagonal pseudocount 0.1, and updates $p \in [10^{-4}, 1 - 10^{-4}]$ from semantic-segment expectations. The left endpoint uses the equilibrium residual lifetime and the right endpoint an analytic survival probability; there is no maximum-duration truncation.

For a component, a Gaussian residual teacher and a neural-logit student share this temporal prior:

$$r(S) \propto P_{B,p}(S) \exp \sum_t e_{t,s_t}, \qquad q_\phi(S \mid X) \propto P_{B,p}(S) \exp \sum_t u_{\phi,t,s_t}, \tag{27}$$

where $e_{t,k} = -\frac{1}{2}[\|x_t - \mu_{t,k}\|^2/(v\tau) + 20 \log(2\pi v\tau)]$. The scalar variance starts at one and is bounded in $[10^{-4}, 10]$; temperature decreases from two at epoch 6 to one at epoch 15. After five

bootstrap epochs with fixed initial-logit responsibilities, each epoch computes teacher responsibilities, updates the decoder by responsibility-weighted branch MSE plus the existing structural regularizers, and trains the student using the teacher marginals. For each target batch $\mathcal{B}$, the student recomputes the partition function over complete components but stops gradients through logits outside $\mathcal{B}$. Its batch objective is $(\log Z(\widetilde{u}) - \sum_{t,k} r_{t,k}\widetilde{u}_{t,k})/|\mathcal{B}|$, where $\widetilde{u}$ retains gradient dependence only at targets in $\mathcal{B}$. The main soft output is the structured student marginal; its hard output is the exact semantic full-path MAP, with deterministic model-ID tie breaking. Teacher outputs are diagnostic. Training receives observations and trajectory identifiers only; full-component inference uses future context.

VIS conditioning and optimization. The diagnostic VIS applies input gain four inside each parent-input function and removes its layerwise spectral normalization (SN):

$$\mu_{k,j} = \alpha_{k,j} f_j\Big(4A^{(k)}_{\mathrm{eff},:,j} \odot x_{t-1}\Big) + \beta_{k,j}. \tag{28}$$

Gain four is fixed as the reciprocal of the common initial selected effective gate, 0.25. Initialization first constructs the canonical SN model, then copies its evaluated effective weights and biases into ordinary linear layers, restoring the random stream consumed by this replacement, before creating the optimizers. The Transformer input is not rescaled, and reported graph scores remain A_{eff}, not $4A_{\mathrm{eff}}$. Hard Top-3 selection, the straight-through rule, structural parameterizations and structural regularizers are retained.

Each model is trained from scratch for 90 epochs: 1,080 decoder and 1,020 student updates using 12 target batches per epoch and disjoint Adam optimizers (learning rate 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, gradient norm clipping at one). Networks remain float32; temporal calculations use float64. CNC and NEC remain zero. The checkpoint minimizes validation unique-target student reconstruction MSE, with the earliest exact tie retained. One selected-checkpoint validation soft-confusion permutation is used for test, graph, teacher, constant-mechanism and supplementary epoch-90 evaluation.

Recovery and fixed-mechanism controls. The structured diagnostic uses graph 11 for protocol development and four reused graph/data/training tuples for evaluation. Across the five tuples, PCDM averages 32.6 uniquely correct private edges and 21.6 all-zero edges. Its paired differences from Independent are -0.0128 $[-0.0474, 0.0218]$ for AP, -0.44 $[-4.24, 3.35]$ percentage points for owner and -0.0064 $[-0.0295, 0.0168]$ for margin. None excludes zero.

Each switching-test trajectory has 23 true boundaries. Matching is one-to-one within two targets. Fixed-mechanism controls use innovation seed $201000+10g+k$ for graph g and mechanism k, with the same generator, zero initialization, 384-step burn-in and 6,144 retained observations. Trained parameters, scaler and validation permutation remain fixed, and no constant-state flag is supplied. Over the 15 controls, PCDM produces 75 false boundaries and Independent 157; the largest error is 53 for Independent on graph 53, mechanism 1. Zero true boundaries make ARI and boundary PRF undefined. The diagnostic changes target aggregation, duration inference and decoder conditioning together, with CNC disabled.

D Observation-Informed Gate Restart

The experiment compares three ways of reassigning the same learned gate values before a common restart. The architecture remains the residual-guided PCDM learner: three states, lag one, hard Top-3, the STE, a width-32 tanh VIS with layerwise spectral normalization and unit input gain, and the Transformer. CNC and NEC are zero. The structured-state diagnostic is a separate comparison and is not a parent for these runs.

D.1 Data, additive screening and the three assignments

The development tuples are $(11, 101, 1001)$, $(23, 103, 1003)$, $(37, 107, 1007)$, $(53, 109, 1009)$ and $(71, 113, 1013)$. The evaluation tuples are $(820001, 820101, 820201)$ through $(820005, 820105, 820205)$, incrementing each coordinate by one. The same additive-sine generator, two CORE parents and one private parent per state, noise SD 0.1, frequency two, 64-step

recurrent schedule and signal-admission rule are used in both blocks. Evaluation graph/data identities and the learning protocol are prespecified. Each graph has one base training stream; evaluation uses independent graphs and observations from the same mechanism family.

Each block uses five selected base models and their additive screening fits. Each graph receives a 90-epoch base model selected by neural-posterior observation-validation MSE. A common cubic-spline basis, defined in Appendix D.6, is fitted to training observations using this model's native neural-state weights. Ridge penalties $\{0.001, 0.1, 10\}$ are compared by observation-validation MSE, with exact ties preferring the larger penalty. This is an additive ridge fit followed by component screening, rather than the sparse-additive optimization algorithm of Ravikumar et al. (2009). The selected candidate is retained without another fit.

The score in Eq. (8) demeans each component over the common training-parent marginal, using the coefficients in their native neural-state order. Neither true states, true supports nor generating-function features enter the fit or reassignment. The validation state permutation is used only for evaluation after model selection.

For each state–target column, let $a_1 \geq \cdots \geq a_{19}$ be its existing non-self gate logits and let $\pi_s(r)$ index the rth largest component score. INFO sets $\Phi^{\mathrm{new}}_{k,\pi_s(r),j} = a_r$. This preserves the exact logit multiset, temperature, diagonal mask and budget. Shared scores, decoder weights, state-affine terms, posterior parameters and buffers initially equal the parent's. KEEP leaves every assignment unchanged. For RANDOM, let m be the overlap of INFO's Top-3 with the parent's Top-3. It samples m parents uniformly from the old three and $3 - m$ from the other 16, then randomizes the ranks within the selected and remaining sets. Thus its replacement count matches INFO separately in every state–target column.

D.2 Common learning schedule and selection

Both Adam states are reset, with learning rate 0.001, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay and separate clipping at one. Shared and sparse penalties retain weight 0.001. The optimizers retain the same parameter groups throughout. During epochs 1–5, structure and posterior gradients are absent; the decoder and affine terms adapt using cached, detached parent training probabilities. The transition matrix is fixed and the variance is reestimated each epoch. Epochs 6–90 restore residual responsibilities, within-window forward–backward, posterior distillation and the transition/variance updates. Teacher temperature reaches one at restart epoch 15 and stays there. The three arms use the same restart random stream and batch order; RANDOM's support draw has its own fixed stream.

Every run completes 90 epochs, comprising 1,080 decoder-optimizer steps and 1,020 posterior steps. Structure parameters receive no update during the first 60 decoder steps. Selection compares neural-q observation-validation MSE at epoch 0 and all 90 subsequent epochs, with the earliest exact tie retained. Epoch 0 is eligible but does not replace full-budget execution. Once selected, a single validation permutation is fixed for all subsequent graph, state, response and constant-mechanism evaluations. The spline is absent from final prediction. The scientific comparison contains 15 development restarts and 20 evaluation-graph trainings (five parents and 15 restarts).

D.3 Primary paired outcomes

Evaluation uses three 6,144-observation realizations per graph, with innovation streams `SeedSequence([820900, graph, replicate, 742])`. Parameters, scalers and validation permutations remain fixed. Each sequence has 95 true boundaries. Three trajectory values are averaged within each graph before pairing INFO with KEEP or RANDOM. The four primary tests receive Holm correction.

Development and evaluation tuples are summarized separately. INFO's development means are PRIVATE AP 0.931 and 54.8 strictly correct owners; its weakest development graph has AP 0.767, 44 correct, 13 all-zero and three wrong owners. On evaluation graphs, mean strict correct and all-zero counts are 58 and two, with no uniquely wrong owners. All five graphs improve on both primary endpoints against both matched controls.

Table 15: Prespecified paired effects. Each block has four primary tests: two endpoints against two controls.

Block	Contrast	Endpoint	Mean difference	Paired 95% CI	Holm p
Eval.	INFO−KEEP	AP	$+0.4319$	$[+0.4116, +0.4521]$	1.95×10^{-6}
Eval.	INFO−KEEP	NMSE	-0.1576	$[-0.1705, -0.1446]$	1.37×10^{-5}
Eval.	INFO−RANDOM	AP	$+0.4382$	$[+0.3782, +0.4981]$	6.96×10^{-5}
Eval.	INFO−RANDOM	NMSE	-0.1625	$[-0.1960, -0.1290]$	1.76×10^{-4}
Dev.	INFO−KEEP	AP	$+0.3563$	$[+0.2642, +0.4484]$	4.26×10^{-4}
Dev.	INFO−KEEP	NMSE	-0.1296	$[-0.1572, -0.1019]$	4.01×10^{-4}
Dev.	INFO−RANDOM	AP	$+0.3832$	$[+0.3151, +0.4513]$	2.94×10^{-4}
Dev.	INFO−RANDOM	NMSE	-0.1581	$[-0.1769, -0.1394]$	7.95×10^{-5}

Intervals use the ordinary paired t distribution with four degrees of freedom. Holm adjusts the four p values within each block, not the intervals. Positive AP and negative NMSE differences favor INFO.

Table 16: Fair owner accuracy for matched restart controls.

Mechanism family	Model	Fair owner accuracy (%) ↑
Additive	KEEP	53.89 ± 3.87
Additive	RANDOM	52.33 ± 3.13
Additive	INFO	$\mathbf{97.78 \pm 3.14}$
Core–private interactions	KEEP	57.11 ± 7.32
Core–private interactions	RANDOM	52.89 ± 6.42
Core–private interactions	INFO	$\mathbf{93.67 \pm 6.18}$

Fair owner accuracy over all 60 private edges. Exact maximum-score ties split one unit of credit equally; an all-zero edge receives one-third credit. Values are mean ± sample SD over five graph/data/training tuples per family and are reported as percentages. Scores use the selected decoder's actual effective graphs and its fixed validation state mapping. Bold marks the highest mean within each family.

D.4 State recovery, fixed mechanisms and support retention

On additive evaluation graphs, INFO's mean boundary F1 is 0.712, compared with 0.533 for KEEP and 0.400 for RANDOM. The paired state-ARI and boundary-F1 intervals against both controls include zero. On development graphs, INFO's boundary F1 is 0.661, below KEEP's 0.681.

Each selected model also reads three constant-mechanism trajectories per graph, using `SeedSequence([820901, graph, state, 743])` on evaluation graphs. Each has 6,143 lag-one targets; the model keeps its three states, scaler, temporal parameters and validation permutation. INFO, KEEP and RANDOM produce respectively 12.1, 54.9 and 84.3 false boundaries per 1,000 targets. Their paired false-boundary-count intervals include zero. Development INFO retains 59.1 false boundaries per 1,000 targets. These rates first average mechanisms within graph; ARI and boundary PRF are undefined for zero true boundaries.

INFO's evaluation AP rises from 0.903 immediately after assignment to 0.975 after learning, and strict correct counts from 52.6 to 58. It retains 870 of 876 initially selected true state–edge coordinates, loses six and gains 16. The following fixed-support comparison tests the role of these later changes.

D.5 A response-amplitude bound for the selected decoder

Let $q_{t,k} \geq 0$ be fixed factual responsibilities with $\sum_k q_{t,k} = 1$. For a selected decoder, write

$$\widehat{m}_j(u; q_t) = \sum_k q_{t,k}\{\alpha_{k,j} f_j(A_{k,:,j} \odot u) + \beta_{k,j}\}, \qquad A_k = A_{\mathrm{eff},k}.$$

With tanh activations and unit input gain, an upper bound on the Lipschitz constant of f_j is $L_j = \prod_\ell \|W^{\mathrm{eff}}_{j,\ell}\|_2$. Here W^{eff} denotes the effective matrix used in the forward pass after spectral normalization (Miyato et al., 2018). All inputs, perturbations and responses use the training-standardized coordinates.

Proposition D.1 (Fixed-parameter response capacity). *For a single-source perturbation $\delta_i e_i$, with the factual responsibilities held fixed,*

$$|\widehat{R}_{t,i,j}| \le C_{t,i,j} := |\delta_i| L_j \sum_k q_{t,k} |\alpha_{k,j} A_{k,i,j}|. \tag{29}$$

For nonnegative evaluation weights w and nonzero true-response energy $T = \|R\|_w^2$,

$$\mathrm{NMSE} \ge B_{\mathrm{capacity}} := \frac{\|(|R| - C)_+\|_w^2}{T} \ge B_{\mathrm{support}}. \tag{30}$$

The statement applies to the given parameters and factual responsibilities.

Proof. The offset $\beta_{k,j}$ cancels in the response difference. The change in the input to f_j has Euclidean norm $|A_{k,i,j}\delta_i|$. Applying its Lipschitz bound to each branch and the triangle inequality to the mixture proves the pointwise upper bound. The reverse triangle inequality gives $|\widehat{R} - R| \ge (|R| - |\widehat{R}|)_+ \ge (|R| - C)_+$. Squaring, weighting and summing gives the first lower bound. If a source–target edge is absent in every state, $C_{t,i,j} = 0$ and its entire true-response energy enters the bound; all other terms are nonnegative. This gives the support lower bound. □

Effective matrix norms are computed in float64 evaluation mode without refreshing spectral-normalization buffers. The saved response banks satisfy the pointwise upper bound. On evaluation graphs, INFO lowers the support/capacity bounds to 0.010/0.029 from KEEP's 0.494/0.585, while actual NMSE remains 0.764. The gap above a bound contains several possible fitting errors; training may change both the bound and the error.

D.6 Conditional functions and fixed informed support

The additive comparisons use the five evaluation tuples and three fresh realizations per graph. Every source uses cubic B-splines with five training-quantile knots, no bias basis and linear extrapolation. Training-parent means and variances standardize the basis across states; columns below variance 10^{-12} are removed. Each state/target has an unpenalized intercept and excludes its own lagged input. For native posterior weights w_{tk}, the fit minimizes

$$\frac{\sum_t w_{tk}(y_{tj} - b_{kj} - z_t^\top \beta_{kj})^2}{\sum_t w_{tk}} + \lambda \|\beta_{kj}\|_2^2.$$

One common $\lambda \in \{0.001, 0.1, 10\}$ per graph/reference minimizes mixture observation-validation MSE, with ties favoring the larger penalty. Selected candidates are retained without refitting.

Screened spline uses the three largest native centered component-RMS scores, the same support as INFO immediately after reassignment, and the base posterior. Support-matched spline instead uses selected INFO support and posterior. Each three-parent fit has 18 basis coefficients plus an intercept per state/target, or 1,140 free coefficients. Thirty candidates fit 1,800 state–target systems. Neither reference learns a state predictor; native component-RMS scores and neural effective-edge scores have different scales.

Fixed-support neural control. INFO-FIXED starts at the INFO reassigned initialization, after rank reassignment and before adaptation. Model parameters, spectral buffers, posterior transition/variance, training responsibilities and random state are identical. The binary mask H^0 is stored throughout learning. With $g = \mathrm{sigmoid}(\Phi/\tau)$ and the original off-diagonal restriction, its gate is

$$G = \left[H^0 + g - \mathrm{stopgrad}(g)\right] \odot g. \tag{31}$$

This retains the original straight-through derivative. Continuous gates, shared scores and neural networks follow the same five-epoch adaptation and 85-epoch joint schedule as INFO; only the forward parent choice is fixed. Each run completes 1,080 mechanism-objective and 1,020 posterior updates, selecting the earliest validation minimum from epochs 0–90. A single selected-model validation permutation is retained for all evaluations. Native support changes are zero at every saved checkpoint.

The paired intervals for the effect of support changes include zero. INFO-FIXED has state ARI 0.941 and boundary F1 0.710, versus INFO's 0.942 and 0.712. Their fixed-mechanism false-boundary rates are 12.63 and 12.13 per 1,000 targets; both have mean mechanism accuracy 0.992. The function-reference contrasts retain a substantial response-fitting gap at the same three-input budget.

Table 17: Learning with fixed or updated informed support on the additive evaluation family. Mean $\pm$ sample SD over five graph/data/training tuples.

Model	PRIVATE AP ↑	Response NMSE ↓	Correct/60 ↑	Owner ↑
INFO-FIXED	0.947 ± 0.073	0.765 ± 0.018	55.2 ± 6.6	0.947 ± 0.073
INFO	$\mathbf{0.975 \pm 0.030}$	$\mathbf{0.764 \pm 0.012}$	$\mathbf{58.0 \pm 2.8}$	$\mathbf{0.978 \pm 0.031}$

Scores are actual neural effective edges. Correct requires a unique correct owner; owner splits credit among exact ties. Response errors average three trajectories within graph. Bold marks the best mean.

Table 18: Prespecified follow-up contrasts: paired mean and ordinary 95% interval, five graphs.

Comparison	Endpoint	Mean [95% interval]	Holm p
INFO − Screened spline	NMSE ↓	$+0.6543$ $[+0.5703, +0.7384]$	2.71e-05
INFO − Support-matched spline	NMSE ↓	$+0.7239$ $[+0.7007, +0.7471]$	2.14e-07
INFO − INFO-FIXED	AP ↑	$+0.0281$ $[-0.0436, +0.0998]$	0.675
INFO − INFO-FIXED	NMSE ↓	-0.0005 $[-0.0103, +0.0092]$	0.884

All differences are neural INFO minus the named comparator. Positive AP favors INFO; positive NMSE favors the comparator. The two function-reference tests and two fixed-support tests are corrected separately. Intervals themselves are not Holm-adjusted.

D.7 Posterior Information for Support Screening

Posterior weighting in screening. NATIVE uses the BASE model's native posterior row for each training observation. POOLED assigns every observation the same vector of state weights, obtained by averaging those rows over training observations. With weight-normalized ridge regularization, its state-specific spline fits coincide. PERMUTED reassigns complete posterior row vectors within each independent training trajectory; state coordinates are preserved. The permutation also changes temporal dependence in the weights, so it does not isolate a single state-misassignment effect.

All conditions share the selected BASE model, cubic spline basis, ridge grid $\{0.001, 0.1, 10\}$, gate-logit multiset, and neural budget. Spline candidates are selected using the normal BASE validation posterior, with exact ties resolved in favor of the larger ridge coefficient. Only screening weights change: neural decoder adaptation, teacher updates and posterior learning use the common protocol. Each restart receives 90 epochs, with five adaptation epochs, and the earliest exact minimum neural validation reconstruction MSE selects among epochs 0–90. The comparison includes ten NATIVE, ten POOLED and ten PERMUTED models; NATIVE models are shared with the restart comparison.

Statistical units. The primary endpoints are PRIVATE AP from the selected neural effective edge scores and response energy NMSE. Three fresh response trajectories are averaged within each graph before matching conditions across five graph/data/training instances per family. Two controls and two endpoints form four prespecified tests per family; Table 19 also retains the joint eight-test Holm sensitivity analysis. These are post hoc screening controls conditional on the observed NATIVE outcomes. Separate significance decisions across mechanism families do not constitute a test of their interaction.

Fixed-mechanism diagnostics. The fixed-mechanism evaluation comprises three state-specific slots for each of thirty models, giving ninety neural-output slots. Ninety additional teacher-output rows are diagnostic views of those slots. These counts are not independent sample sizes. Neural outputs and teacher diagnostics are summarized separately after averaging the three state-specific slots within each graph.

Table 19: Paired effects of native posterior screening. Differences are NATIVE minus the indicated control.

Control	Endpoint	Difference [95% CI]	Raw p	Holm (4)	Holm (8)
Additive					
POOLED	PRIVATE AP	$+0.1953$ $[+0.0695, +0.3211]$	0.0126	0.0251	0.0753
POOLED	Response NMSE	-0.0385 $[-0.0641, -0.0128]$	0.0140	0.0251	0.0753
PERMUTED	PRIVATE AP	$+0.2489$ $[+0.1334, +0.3643]$	0.0039	0.0157	0.0313
PERMUTED	Response NMSE	-0.0422 $[-0.0660, -0.0183]$	0.0080	0.0239	0.0559
Interactions					
POOLED	PRIVATE AP	$+0.1491$ $[+0.0501, +0.2481]$	0.0139	0.0553	0.0753
POOLED	Response NMSE	-0.0155 $[-0.0258, -0.0052]$	0.0138	0.0553	0.0753
PERMUTED	PRIVATE AP	$+0.1141$ $[+0.0337, +0.1946]$	0.0170	0.0553	0.0753
PERMUTED	Response NMSE	-0.0099 $[-0.0210, +0.0011]$	0.0668	0.0668	0.0753

Five paired graph-level differences per family; ordinary two-sided t_4 intervals are not multiplicity-adjusted. Holm (4) is the prespecified family of two controls by two endpoints within each mechanism family. Holm (8) jointly adjusts all eight tests as a sensitivity analysis. Positive AP and negative NMSE differences favor NATIVE.

Table 20: Fixed-mechanism diagnostics for posterior screening.

Mechanisms	Screening	Neural accuracy ↑	Neural false boundaries per 1,000 ↓	Teacher false boundaries per 1,000 ↓
Additive	NATIVE	$\mathbf{0.992 \pm 0.005}$	$\mathbf{12.13 \pm 6.09}$	$\mathbf{4.70 \pm 3.11}$
	POOLED	0.928 ± 0.143	35.01 ± 52.56	46.60 ± 94.04
	PERMUTED	0.928 ± 0.143	41.26 ± 62.51	48.29 ± 93.91
Interactions	NATIVE	$\mathbf{0.925 \pm 0.114}$	$\mathbf{81.60 \pm 99.78}$	$\mathbf{72.79 \pm 108.24}$
	POOLED	0.867 ± 0.133	143.15 ± 133.11	132.78 ± 131.23
	PERMUTED	0.904 ± 0.118	110.66 ± 118.68	103.46 ± 120.33

Mean $\pm$ sample SD across five graph-level averages; each average uses three fixed-state slots. Accuracy refers to neural mechanism assignment. Teacher outputs are separate diagnostics of the same slots. True boundary counts are zero. Bold marks the best mean within each family and column.

D.8 Complete-Candidate Ranking and Learning Stages

The complete edge–state evaluation contains 120 CORE-active, 60 PRIVATE-active, 120 PRIVATE-inactive and 840 ABSENT coordinates. It retains each selected model's validation state mapping. Table 21 compares selected models; the supplementary ALL-STATE tests in Table 22 retain separate four-test correction families for the restart controls and the screening-weight controls. Tables 23 and 24 report the saved learning stages using the same effective-score definition.

Table 21: Structure ranking over the complete edge–state candidate set and the corresponding hard support.

Condition	ALL-STATE AP ↑	PRIVATE AP ↑	UNION AP ↑	Mean TP/FP/FN
Additive				
KEEP	0.469 ± 0.009	0.543 ± 0.023	0.517 ± 0.016	79.4/100.6/100.6
RANDOM	0.493 ± 0.030	0.537 ± 0.022	0.551 ± 0.023	86.8/93.2/93.2
INFO	$\mathbf{0.981 \pm 0.020}$	$\mathbf{0.975 \pm 0.030}$	$\mathbf{0.978 \pm 0.032}$	$\mathbf{177.2/2.8/2.8}$
POOLED	0.861 ± 0.086	0.779 ± 0.130	0.873 ± 0.044	155.8/24.2/24.2
PERMUTED	0.840 ± 0.079	0.726 ± 0.119	0.850 ± 0.047	152.6/27.4/27.4
Interactions				
KEEP	0.547 ± 0.029	0.584 ± 0.070	0.616 ± 0.016	91.8/88.2/88.2
RANDOM	0.519 ± 0.053	0.535 ± 0.058	0.589 ± 0.040	89.0/91.0/91.0
INFO	$\mathbf{0.922 \pm 0.048}$	$\mathbf{0.939 \pm 0.058}$	$\mathbf{0.960 \pm 0.022}$	$\mathbf{165.8/14.2/14.2}$
POOLED	0.861 ± 0.068	0.790 ± 0.125	0.904 ± 0.032	154.8/25.2/25.2
PERMUTED	0.879 ± 0.059	0.825 ± 0.093	0.906 ± 0.040	158.4/21.6/21.6

AP values are mean ± sample SD across five graphs per family. ALL-STATE AP uses actual decoder effective scores on 1,140 directed nonself edge–state coordinates (180 active, 960 inactive). PRIVATE AP uses 180 private-edge coordinates; UNION AP uses 380 directed edges after maximizing across states. Hard counts use the actual 180-coordinate support. INFO is NATIVE in the posterior-screening comparison. Bold marks the highest mean within each family and metric.

Table 22: Supplementary paired ALL-STATE AP differences, INFO minus each control.

Mechanisms	Control	Difference [95% CI]	Raw p	Holm (4)
KEEP/RANDOM control comparisons				
Additive	KEEP	$+0.5120$ $[+0.4828, +0.5413]$	1.07e-06	4.27e-06
Additive	RANDOM	$+0.4887$ $[+0.4366, +0.5409]$	1.3e-05	3.89e-05
Interactions	KEEP	$+0.3758$ $[+0.3050, +0.4466]$	0.000123	0.000246
Interactions	RANDOM	$+0.4038$ $[+0.2992, +0.5084]$	0.000429	0.000429
Posterior-screening supplementary comparisons				
Additive	POOLED	$+0.1208$ $[+0.0386, +0.2029]$	0.0151	0.0452
Additive	PERMUTED	$+0.1412$ $[+0.0671, +0.2153]$	0.00612	0.0245
Interactions	POOLED	$+0.0618$ $[+0.0161, +0.1074]$	0.0198	0.0452
Interactions	PERMUTED	$+0.0439$ $[-0.0034, +0.0912]$	0.0615	0.0615

Five paired graph-level differences per mechanism family; ordinary t_4 intervals are not multiplicity-adjusted. Each four-row block forms its own Holm family across the two mechanism families. These post hoc analyses are separate from the primary PRIVATE AP/response NMSE tests and the posterior-screening primary eight tests. INFO denotes NATIVE in the lower block.

Table 23: Mean ALL-STATE AP at saved learning stages.

Condition	BASE	0	5	30	60	90	Selected
Additive							
KEEP	**0.314**	0.314	0.314	0.342	0.405	0.467	0.469
RANDOM	**0.314**	0.276	0.276	0.330	0.396	0.488	0.493
INFO	**0.314**	**0.967**	**0.967**	**0.968**	**0.974**	**0.979**	**0.981**
POOLED	0.301	0.710	0.710	0.836	0.851	0.862	0.861
PERMUTED	0.301	0.704	0.704	0.802	0.832	0.833	0.840
Interactions							
KEEP	0.396	0.396	0.396	0.447	0.495	0.542	0.547
RANDOM	0.396	0.295	0.295	0.351	0.424	0.509	0.519
INFO	0.396	**0.947**	**0.947**	**0.879**	**0.898**	**0.922**	**0.922**
POOLED	**0.401**	0.708	0.708	0.853	0.845	0.852	0.861
PERMUTED	0.396	0.714	0.714	0.858	0.865	0.883	0.879

Each AP entry averages five graphs. All nodes use each model's selected-checkpoint validation mapping, so BASE values can differ between conditions. BASE is before reassignment; 0 is immediately after it. Selected uses each model's actual validation-selected epoch. Screening component-RMS scores are a separate object. Bold marks the best mean among conditions within each mechanism family and stage, not across epochs.

Table 24: True hard-support coordinates at saved learning stages, summed over five graphs.

Condition	BASE	0	5	30	60	90	Selected
Additive							
KEEP	**323**	323	323	320	350	389	397
RANDOM	**323**	253	253	303	357	427	434
INFO	**323**	**876**	**876**	**877**	**881**	**884**	**886**
POOLED	317	696	696	774	775	781	779
PERMUTED	317	689	689	746	761	757	763
Interactions							
KEEP	**362**	362	362	387	415	454	459
RANDOM	**362**	272	272	300	361	436	445
INFO	**362**	**861**	**861**	**803**	**814**	**829**	**829**
POOLED	**362**	700	700	774	761	766	774
PERMUTED	**362**	706	706	784	782	796	792

Each model selects 180 coordinates and each true graph contains 180 active coordinates. Across five graphs, FP and FN both equal 900 minus the displayed TP. Interaction-family INFO loses true support from 861 at epoch 0 to 829 at selection, while FP rises from 39 to 71. Bold marks the best mean among conditions within each mechanism family and stage.

D.9 CORE–PRIVATE INTERACTION MECHANISMS

Generation and calibration. For target j, let c_j be the first CORE parent sampled by the fixed graph-generation rule, and p_{kj} its PRIVATE parent in state k. The conditional mean is

$$f_{k,j}(x) = \sum_i W_{kij} \sin(2x_i) + \gamma_{kj} \sin(2x_{c_j}) \sin(2x_{p_{kj}}), \tag{32}$$

$$\gamma_{kj} = 0.5 \ \mathrm{sign}(W_{k,p_{kj},j}) \min\{|W_{k,p_{kj},j}|, |W_{k,c_j,j}|\}. \tag{33}$$

Each state retains two CORE parents and one PRIVATE parent per target, with noise SD 0.1. The five graph/data/training tuples are $(830001, 830101, 830201)$ through $(830005, 830105, 830205)$. Neither true states nor generating functions enter model training or selection.

Bounded response-strength calibration starts from the initial proposal W_0. It updates 100 union-edge magnitudes per graph; each CORE magnitude is shared across all three states. Signs, supports, interaction coefficient $\rho = 0.5$ and noise SD $\sigma = 0.1$ remain fixed. Starting magnitudes $a_{e,0} \in [0.58, 0.62]$ are constrained to $[0.9a_{e,0}, 1.5a_{e,0}]$. For the smallest active-state median response/noise ratio b_e of edge e, an update is

$$a_e^+ = \mathrm{clip}_{[0.9a_{e,0}, 1.5a_{e,0}]} \left[a_e \, \mathrm{clip}_{[0.9,1.1]} \left(\sqrt{2/\max(b_e, 10^{-12})} \right) \right].$$

The calibration innovation stream is fixed; interaction coefficients are recomputed after each update. Each graph permits at most 20 amplitude updates and stops at its first complete gate pass. All five graphs are then frozen before one independent audit per graph, using the frozen calibration perturbation scale without feeding audit outcomes back into calibration. Every active coordinate must exceed 1.5 noise SD; CORE and PRIVATE class medians must lie in $[1.5, 2.5]$, and their ratio in $[0.75, 1.3334]$. These rules use the full function, including its product term. The five graphs require one or two updates. Interaction-difference energy is 7.65–8.23% of full response energy at calibration scale and 7.60–8.18% at the evaluation scale. These are bounded interactions; the coefficient 0.5 is not an energy fraction. Because response-strength calibration also changes coefficient magnitudes, the difference from the additive-family scores is not an isolated interaction effect.

Learning and evaluation. Each 6,144-observation formal trajectory uses TRAIN $[0, 3072)$ and validation $[3072, 4608)$, with a TRAIN-only scaler; its remaining tail is unused. Lag-one windows have length 32 and strides 16/32. Every base model trains from scratch. INFO, KEEP and RANDOM inherit the same selected base and follow Appendix D, including the same within-tuple training stream. All 20 runs complete 90 epochs, 1,080 decoder and 1,020 posterior updates. Bases select from epochs 1–90; restarts select from epochs 0–90. The earliest neural-q validation MSE minimum is retained. S1 uses exactly three ridge candidates per graph, with no additional refit.

All neural and S1 selections precede evaluation. One validation soft-confusion permutation is then fixed for every model's graphs and state outputs. Each graph has three independent 6,144-observation realizations, using `SeedSequence([830900, graph, replicate, 742])`. Finite responses follow Appendix D.12, with the product term included in the true response. Calibration and evaluation perturbations are respectively 0.1 and 0.008 of their prescribed source spans. Trajectories are averaged within graph before pairing across the five graphs. The four primary tests form one Holm family, separate from the additive family; ordinary t intervals have four degrees of freedom.

State errors and support changes. INFO improves both primary endpoints on all five graphs against both controls. It leaves 3.6 all-zero and 1.4 uniquely wrong private edges on average. State recovery is less uniform: graph 830003 has ARI 0.484 and boundary F1 0.101, while graphs 830002 and 830005 have lower boundary F1 than KEEP. The paired ARI and boundary-F1 intervals against both controls include zero. At selection, INFO retains 165.8 true and 14.2 false state–edge coordinates per graph on average.

From the informed initialization to the selected neural checkpoint, INFO's true hard-support coordinates fall from 861 to 829 across five graphs: 27 are gained and 59 lost. False coordinates rise from 39 to 71. Over the same interval, PRIVATE AP rises from 0.856 to 0.939 and uniquely correct owners from 50 to 55 per graph. Improved effective-edge ranking therefore coexists with a net loss of true hard-support coordinates.

For each graph, three constant-mechanism observations use `SeedSequence([830901, graph, state, 743])`. Models retain their three states, scaler, temporal parameters and validation permutation. No constant-state indicator is supplied. There are 60 model–trajectory evaluations with 6,143 valid targets each. INFO averages 81.60 false boundaries per 1,000 targets, compared with 201.37 for KEEP and 200.38 for RANDOM. The corresponding paired differences in false boundaries per trajectory are -735.7 $[-1745.6, 274.1]$ and -729.7 $[-1507.4, 48.1]$; both intervals include zero. All models remain oversegmented under fixed mechanisms. On graph 830003, mechanism 0 alone produces 3,190 false boundaries for INFO, with accuracy 0.310. Zero true boundaries make ARI and boundary PRF undefined.

Conditional reference. The all-source additive reference has response NMSE 0.205 ± 0.064, compared with neural INFO's 0.823 ± 0.032. Its component-RMS AP is 0.922 ± 0.087; all five fits select ridge weight 0.001. It uses all 19 candidate inputs, an additive spline family and the base model's fixed posterior. It is therefore a broader conditional fit, not a comparison with INFO's final three inputs and posterior held equal. Its component-RMS scores are reported separately from neural effective-edge scores.

D.10 Function fitting with the same support and posterior

We fix each INFO model's hard three-parent support and native posterior, and fit an additive cubic-spline reference to the same observed training data. The reference uses five training-quantile knots, degree three, linear extrapolation, training-only centering and scaling, an unpenalized intercept, and posterior-weighted ridge normalized by the state weight sum. Validation reconstruction MSE selects the penalty from $\{0.001, 0.1, 10\}$, with exact ties resolved toward the larger penalty; all five graphs select 0.001. Each selected fit is used directly. Both functions use the same response inputs, perturbations and fixed factual posterior.

The spline reference lowers response NMSE on all five graphs (Table 25). The mean reduction is 0.622, with paired 95% interval $[0.548, 0.695]$ and $p = 1.95 \times 10^{-5}$. Reconstruction MSE also decreases. This comparison isolates a substantial conditional function-fitting gap after the inputs and posterior have been learned. Function class, regularization and solver change together; state and support measurements remain those of the frozen INFO provider.

Table 25: Conditional function comparison with the same INFO support and posterior. Each graph averages three fresh trajectories; mean and sample SD use the five graph/data/training blocks. The paired test is a standalone comparison with four degrees of freedom.

	Response NMSE ↓		Reconstruction MSE ↓	
Graph	Neural INFO	Fixed-H/q spline	Neural INFO	Fixed-H/q spline
830001	0.815	**0.192**	0.494	**0.130**
830002	0.792	**0.161**	0.520	**0.108**
830003	0.853	**0.317**	0.539	**0.237**
830004	0.794	**0.177**	0.502	**0.133**
830005	0.859	**0.156**	0.533	**0.108**
Mean ± SD	0.823 ± 0.032	$\mathbf{0.201 \pm 0.067}$	0.517 ± 0.020	$\mathbf{0.143 \pm 0.054}$

The response RMS ratio increases from 0.404 to 0.896, and energy alignment increases from 0.421 to 0.893. Graph 830003 retains the largest error. Error on inactive coordinates, normalized by total true response energy, increases from 0.00345 to 0.01390; the response improvement therefore does not eliminate leakage onto inactive responses.

D.11 Candidate-Score Tying at Fixed Screening Information

Comparison. We compare tied and untied candidate scores on the five additive graph/data/training tuples $(820001, 820101, 820201)$ through $(820005, 820105, 820205)$ at regime lengths $L = 64$ and $L = 32$. Within each length and tuple, both arms start from the same base model, native inferred posterior and INFO gate assignment. The untied arm copies the shared raw candidate logits into three independently learnable matrices, adding 800 parameters. Initial effective graphs and decoder

outputs are identical. State-local gates, the hard three-parent budget, STE, shared VIS functions and state-specific affine terms are retained. Thus the intervention unties candidate scores while keeping the neural mechanism functions shared.

Each length uses its own base posterior and observation data. The posterior is frozen throughout all 90 conditional-fitting epochs. A fresh Adam optimizer uses learning rate 0.001, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and gradient clipping at one. The first five epochs adapt the decoder with structure fixed; the remaining 85 update the candidate scores, gates and decoder. The objective is branch reconstruction error weighted by the fixed native posterior, with the original affine and effective-edge sparsity penalties, both weighted by 0.001. Each run has 1,080 mechanism updates and zero posterior updates. The earliest exact minimum validation mixture MSE over epochs 0–90 selects the model. There are 20 conditional models in total.

Evaluation and statistics. Each model is evaluated on three fresh observations using the fixed provider posterior and validation state permutation. PRIVATE AP uses actual decoder effective scores; finite-response NMSE follows Equation (9). We first average trajectories within each graph, then compute paired differences over five graphs. The four length-specific comparisons form the primary Holm family. Two difference-in-differences tests form a separate interaction family; a six-test Holm correction is also reported. Ordinary t intervals have four degrees of freedom and are not multiplicity-adjusted. Differences are tied minus untied; length interactions subtract the $L = 64$ difference from the $L = 32$ difference.

Table 26: Paired candidate-score effects and regime-length interactions.

Contrast	Endpoint	Difference [95% CI]	Raw p	Holm p	Six-test p
$L = 64$	AP	$+0.00113$ $[-0.04572, +0.04797]$	0.9500	1.0000	1.0000
$L = 64$	NMSE	-0.01081 $[-0.01675, -0.00486]$	0.0072	0.0290	0.0435
$L = 32$	AP	-0.00543 $[-0.03838, +0.02752]$	0.6710	1.0000	1.0000
$L = 32$	NMSE	-0.00005 $[-0.01049, +0.01039]$	0.9901	1.0000	1.0000
$\Delta_{32} - \Delta_{64}$	AP	-0.00656 $[-0.06589, +0.05278]$	0.7744	0.7744	1.0000
$\Delta_{32} - \Delta_{64}$	NMSE	$+0.01076$ $[-0.00349, +0.02501]$	0.1040	0.2081	0.5202

Δ_L is Tied minus Untied at length L. Intervals use the five paired tuples and t_4. The four within-length tests and two interaction tests have separate Holm families; the last column adjusts all six together.

The PRIVATE AP intervals include zero at both lengths. At $L = 64$, tying lowers response NMSE by 0.01081, with a six-test Holm-adjusted $p = 0.0435$. The $L = 32$ response difference and both length interactions are inconclusive. The response reduction is an absolute NMSE difference. These results concern candidate-score tying under fixed inferred states, rather than a comparison of fully independent state networks.

Fixed-mechanism observations. Both arms inherit the same state predictions. Across three fixed-mechanism trajectories per graph, the frozen posterior produces 123.51 false boundaries per 1,000 targets at $L = 64$ and 207.89 at $L = 32$; corresponding mean mechanism accuracies are 0.836 and 0.722. Each length has 15 distinct observations evaluated by both arms. With no true boundaries, boundary precision, recall and F1 are undefined. Candidate-score tying does not alter this oversegmentation because the posterior is fixed.

D.12 Response sampling and complete learning costs

Each fresh trajectory starts at zero, discards 384 burn-in observations and retains 6,144. State and reconstruction metrics use all 6,143 lag-one targets. Responses use 48 uniformly spaced nonoverlapping 32-target windows, all 20 sources and both signs. Perturbations are 0.008 of the window's source $Q_{0.95} - Q_{0.05}$ span. Factual probabilities remain fixed during decoder perturbation, as do trained parameters, scalers and validation permutations across trajectories. Development innovation streams use `SeedSequence([20260921, graph, replicate, 741])`; evaluation streams are specified with each mechanism family. Trajectory values are averaged within graph before paired comparisons.

Table 27: Computational components of INFO over five tuples per family. Times are mean ± sample SD.

Mechanism family	Neural stages	Neural intervals (min)	Screening CPU (s)
Additive	$90 + 90$ epochs	124.18 ± 6.21	0.85 ± 0.06
Interactions	$90 + 90$ epochs	92.79 ± 1.92	0.82 ± 0.01

Each complete INFO procedure includes base learning, three screening candidates and neural restart, totaling 2,160 mechanism-objective and 2,040 posterior updates. Neural intervals sum the two recorded process durations on a shared GPU; screening uses CPU process time and is reported separately. These are component timings, excluding scheduling and other end-to-end overhead.

Conditional references also depend on the trained posterior and, when support-matched, the selected INFO support. The interaction-family fixed-H/q spline fits 15 candidates comprising 900 state–target systems, with five selected fits of 1,140 coefficients each. Fitting and fresh-trajectory scoring take 26.1 CPU seconds. Producing their support and posterior additionally requires base learning, screening and INFO restart: the measured components total 24,240.1 CPU seconds and 27,759.9 neural wall seconds. The latter excludes screening wall time. These are component costs on a shared workstation.

D.13 Computational scaling

Let $B, W, d, K, h, h_{\mathrm{ff}}$ denote batch size, window length, variable count, state count, hidden width and feed-forward width. Posterior attention costs $O(BW^2h + BW(h^2 + hh_{\mathrm{ff}}))$. Shared scores and state gates store $(K+1)d^2$ structural parameters before diagonal exclusions.

Two-layer VIS dynamics cost $O(BWKd(dh + h^2))$ using dense masked-input kernels. CNC adds a perturbed-window VIS evaluation; NEC adds approximately $(E' + 1)H$ short-step predictions. Sequence length increases the number of windows; variable count increases structural storage and decoder work.

Table 28 gives full-epoch wall times and peak allocated memory for five input sizes. Each measurement uses three warm-up updates followed by parameter, buffer and RNG restoration, then one complete non-bootstrap epoch including teacher inference, both optimizers, validation and checkpoint I/O. The settings are $K = 3$, window 32, stride 16, batch 16 and hidden width 32. Results average three timing sessions per case on an RTX 6000D and Ryzen 9 9950X3D with PyTorch 2.11.0, float32 and two CPU threads. The GPU is shared, so these are observed workstation costs. Top-k uses dense masked-input kernels and does not reduce operator dimensions.

Table 28: Full-epoch costs on synthetic inputs, with a shared GPU. Wall time is mean ± SD over three sessions; memory is the mean per-process peak.

Variables	Train length	Condition	Wall time (s) ↓	Allocated memory (MiB) ↓
20	3072	CNC-OFF	$\mathbf{42.0 \pm 0.3}$	**77.2**
20	3072	CNC-ON	52.8 ± 0.8	85.3
20	3072	CNC-FG	53.5 ± 0.7	85.6
50	3072	CNC-OFF	$\mathbf{67.5 \pm 0.1}$	**102.1**
50	3072	CNC-ON	93.4 ± 0.4	130.7
50	3072	CNC-FG	93.6 ± 0.4	130.8
100	3072	CNC-OFF	$\mathbf{126.5 \pm 0.8}$	**172.5**
100	3072	CNC-ON	174.9 ± 0.7	256.6
100	3072	CNC-FG	176.6 ± 2.1	260.6
20	6144	CNC-OFF	$\mathbf{84.6 \pm 0.4}$	**84.1**
20	6144	CNC-ON	107.9 ± 0.4	88.0
20	6144	CNC-FG	107.3 ± 0.3	87.2
20	12288	CNC-OFF	$\mathbf{167.3 \pm 0.7}$	97.9
20	12288	CNC-ON	215.3 ± 0.9	97.7
20	12288	CNC-FG	215.4 ± 0.5	**97.3**

CNC-OFF disables the regularizer; CNC-ON holds the denominator fixed; CNC-FG differentiates it. Both use weight 0.001. Measurements include validation and checkpoint I/O under shared-device contention.

E Supplementary Ablations and Diagnostic Controls

E.1 Support-rewiring controls

The simulator uses 20 variables, three recurring states and 6,144 observations:

$$\mathbf{x}_t = \tanh(\mathbf{x}_{t-1} W_{S_t}) + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, 0.1^2 I).$$

Each state has three non-self parents per target. Weight columns have absolute sums at most 0.64; signal and contraction parameters are 0.8 and state modulation is 0.15. After 128 burn-in steps, states cycle with equal occupancy. Diffuse rewiring spreads changes over targets; concentrated rewiring changes selected targets. The five graph/data/training tuples are $(11, 101, 1001)$, $(23, 103, 1003)$, $(37, 107, 1007)$, $(53, 109, 1009)$ and $(71, 113, 1013)$.

Train/validation/test fractions are 50/25/25%, with windows confined to their split. Window length is 32, strides are 16/32, and duplicate lag-one predictions are averaged. Adam uses learning rate 10^{-3}, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, batch 16, zero weight decay and clipping at one. Validation MSE selects within each stated budget. The posterior has one Transformer layer, width 32, four heads, feed-forward width 64 and no dropout. VIS has two width-32 tanh layers with spectral normalization. Single-graph uses one state; Shared-only fixes gates to one; PCDM learns shared scores and state gates; Independent learns separate scores with their mean used for structural regularization.

Union AP ranks $\max_k (A_k)_{ij}$ over 380 directed non-self candidates, grouping exact ties. F1 uses threshold 0.1. A three-state Top-3 model permits union indegree nine. Sample SD and paired exploratory t intervals use five tuples.

Table 29: Support-rewiring controls at $r = 0.2$, $L = 64$ (mean $\pm$ SD). Column groups distinguish the auxiliary settings. LOCAL F1 here differs from PRIVATE edge–state AP.

	CNC/NEC off		CNC/NEC enabled		
Model	Union AP ↑	Union F1 ↑	CORE AP ↑	LOCAL F1 ↑	ABSENT FPR ↓
M0: single	$\mathbf{0.6009 \pm 0.0257}$	$\mathbf{0.5889 \pm 0.0320}$	0.359 ± 0.046	$\mathbf{0.272 \pm 0.035}$	0.113 ± 0.011
M1: shared only	0.5602 ± 0.0365	0.5361 ± 0.0525	0.343 ± 0.036	0.266 ± 0.044	$\mathbf{0.112 \pm 0.016}$
M2: PCDM	0.5655 ± 0.0437	0.4518 ± 0.0378	$\mathbf{0.409 \pm 0.029}$	0.193 ± 0.052	0.272 ± 0.021
M3: independent	0.5703 ± 0.0357	0.4464 ± 0.0349	–	0.247 ± 0.074	0.265 ± 0.018

CORE: state-support intersection; LOCAL: union minus CORE; ABSENT: off-diagonal complement. CORE AP ranks the shared factor (absent in M3). Union AP ranks continuous effective-union scores. At threshold 0.1, Union F1 and ABSENT FPR use the union graph, while LOCAL F1 pools aligned state–LOCAL-edge coordinates.

At $r = 0.2$, the graphs contain 48 CORE, 36 LOCAL and 296 ABSENT candidate edges. Active and inactive LOCAL edges overlap substantially in score. Table 30 gives the six rewiring levels and their measured support intersections and unions.

Table 30: Complete diffuse-rewiring comparison at $L = 64$: effective-union AP (mean $\pm$ sample SD over five tuples).

r	$\lvert \cap E_k \rvert / \lvert \cup E_k \rvert$	Ratio	M0 ↑	M1 ↑	M2 ↑	M3 ↑
0	60/60	1.0000	$\mathbf{0.4136 \pm 0.0544}$	0.3928 ± 0.0858	0.3585 ± 0.0699	0.2615 ± 0.0311
0.1	54/72	0.7500	0.3965 ± 0.0558	$\mathbf{0.4022 \pm 0.0460}$	0.3256 ± 0.0564	0.2936 ± 0.0442
0.2	48/84	0.5714	0.3867 ± 0.0188	$\mathbf{0.3971 \pm 0.0443}$	0.3511 ± 0.0192	0.3162 ± 0.0404
0.3	42/96	0.4375	$\mathbf{0.3925 \pm 0.0198}$	0.3882 ± 0.0528	0.3679 ± 0.0267	0.3151 ± 0.0355
0.4	36/108	0.3333	0.4071 ± 0.0339	$\mathbf{0.4244 \pm 0.0403}$	0.3694 ± 0.0230	0.3362 ± 0.0406
0.6	24/132	0.1818	$\mathbf{0.4418 \pm 0.0338}$	0.4368 ± 0.0274	0.4052 ± 0.0349	0.3796 ± 0.0150

M0: Single-graph; M1: Shared-only; M2: PCDM; M3: Independent. All rows use direct mixture reconstruction, 30 epochs, 50/25/25% splits, $\lambda_{\mathrm{CNC}} = 0.001$ and $\lambda_{\mathrm{NEC}} = 0.03$. This is the auxiliary-enabled scan; the central auxiliary-off comparison is in Table 29. Bold identifies the largest mean among the four methods at each rewiring level. Ratios are measured support counts, not nominal $1 - r$.

Training settings and auxiliary losses. Direct mixture reconstruction uses shared-affine, effective-edge sparsity and smoothness weights $0.001, 0.001, 0.01$, with DAG weight zero. Entropy $-\langle \sum_k q_k \log q_k \rangle$ and marginal-balance $\sum_k \bar{q}_k \log(K\bar{q}_k)$ penalties each have weight 0.001; smoothness is the mean squared posterior change over valid adjacent pairs. CNC and NEC use the settings below when enabled. Additional free-running reconstruction and oracle-supervision weights are zero. Static comparisons may also use $L_{\mathrm{DAG}} = [\mathrm{tr}\exp(A \circ A) - d]^2$ with the weights given alongside each benchmark (Zheng et al., 2018).

At the central 30-epoch condition, union AP is 0.5655 with both auxiliaries off, 0.5761 with CNC alone, 0.2721 with NEC alone and 0.3511 with both. All tested positive NEC weights $\{0.003, 0.01, 0.03\}$ reduce AP; increasing the modeled state count to five also lowers AP and increases fragmentation. State ARI stays near zero across the CNC grid. NEC is therefore disabled in the primary residual-guided learner.

With both auxiliaries enabled, PCDM-minus-Independent AP intervals include zero for diffuse $r = 0.2$ rewiring at segment lengths 32, 64 and 256; at length 32, all five F1 differences are negative. Concentrating rewiring gives AP differences 0.0506 $[0.0064, 0.0949]$ at $r = 0.2$ and 0.0437 $[0.0143, 0.0731]$ at $r = 0.4$, while the latter F1 interval includes zero. Given-state responsibilities raise union AP to 0.4873 ± 0.0457 for PCDM and 0.4322 ± 0.0117 for Independent, with paired difference 0.0551 $[0.0046, 0.1055]$; both worsen reconstruction in every tuple. At 90 epochs, PCDM's intervals against Shared-only and Independent include zero. These state-information and budget controls preserve the same observations.

Tables 31 and 32 give the corresponding capacity, sparsity and auxiliary-loss sensitivities under direct reconstruction.

Table 31: Direct-reconstruction sensitivity (mean ± sample SD over five tuples).

(a) State capacity, indegree and sparsity

Parameter	Value	Union AP ↑	Union F1 ↑	Recon. MSE ↓	Boundaries
Default	–	0.3511 ± 0.0192	0.3396 ± 0.0295	0.8882 ± 0.0133	910.0 ± 28.1
K	2	$\mathbf{0.3920 \pm 0.0407}$	$\mathbf{0.3654 \pm 0.0356}$	–	665.8 ± 36.9
K	5	0.2454 ± 0.0246	0.2861 ± 0.0530	–	1177.6 ± 21.9
$d_{\max}$	1	0.2698 ± 0.0287	0.1676 ± 0.0462	0.9172 ± 0.0149	–
$d_{\max}$	5	0.2418 ± 0.0071	0.3014 ± 0.0074	$\mathbf{0.8815 \pm 0.0103}$	–
$\lambda_{\mathrm{sparse}}$	0	0.3426 ± 0.0239	0.3419 ± 0.0361	0.8887 ± 0.0131	–
$\lambda_{\mathrm{sparse}}$	0.01	0.3699 ± 0.0412	0.3593 ± 0.0338	0.8857 ± 0.0131	–

(b) NEC strength, candidate draws and horizon

Parameter	Value	Union AP ↑	Union F1 ↑	Recon. MSE ↓
Default	–	0.3511 ± 0.0192	0.3396 ± 0.0295	0.8882 ± 0.0133
λ_{NEC}	0	$\mathbf{0.5761 \pm 0.0507}$	$\mathbf{0.4519 \pm 0.0450}$	$\mathbf{0.8793 \pm 0.0147}$
λ_{NEC}	0.003	0.4343 ± 0.0305	0.3718 ± 0.0388	0.8842 ± 0.0147
λ_{NEC}	0.01	0.4040 ± 0.0473	0.3625 ± 0.0472	0.8838 ± 0.0133
H	1	0.3396 ± 0.0522	0.3290 ± 0.0425	0.8869 ± 0.0163
E	2	0.5668 ± 0.0555	0.4423 ± 0.0622	0.8800 ± 0.0169
H	4	0.3448 ± 0.0413	0.3331 ± 0.0356	0.8899 ± 0.0119

The common setting is $K = 3$, indegree 3, sparsity 0.001, CNC 0.001, NEC 0.03, $E = 4$, $H = 2$ and 30 epochs on the $r = 0.2, L = 64$ tanh controls. Each row changes only its named setting. Dashes denote unreported measurements. These are direct-reconstruction sensitivities.

Table 32: CNC weight and perturbation scan under direct reconstruction. All cells use five tuples; entries are mean ± sample SD.

(a) Structural ranking and state agreement

λ	δ	Union AP ↑	ARI ↑
0.0002	0.001	0.3062 ± 0.0321	0.00017 ± 0.00051
0.0002	0.008	0.3128 ± 0.0260	0.00058 ± 0.00033
0.0002	0.032	0.2967 ± 0.0231	0.00021 ± 0.00023
0.001	0.001	0.3320 ± 0.0279	0.00054 ± 0.00074
0.001	0.008	0.3511 ± 0.0192	-0.00010 ± 0.00033
0.001	0.032	0.3467 ± 0.0476	0.00080 ± 0.00072
0.005	0.001	0.3594 ± 0.0260	$\mathbf{0.00086 \pm 0.00050}$
0.005	0.008	$\mathbf{0.3839 \pm 0.0283}$	0.00047 ± 0.00125
0.005	0.032	0.3749 ± 0.0260	0.00058 ± 0.00142

(b) Response variation and floor frequency

λ	δ	Within-state MSE	Between-state MSE	Floor rate
0.0002	0.001	1.1062 ± 0.2079	1.1264 ± 0.4843	0.0625 ± 0.0225
0.0002	0.008	0.9971 ± 0.1476	1.0002 ± 0.5481	0.0430 ± 0.0473
0.0002	0.032	0.9639 ± 0.2343	0.9540 ± 0.5299	0.0309 ± 0.0278
0.001	0.001	0.5862 ± 0.1572	0.5728 ± 0.2694	0.2080 ± 0.0344
0.001	0.008	0.6883 ± 0.1565	0.5318 ± 0.1485	0.0306 ± 0.0274
0.001	0.032	0.7957 ± 0.1501	0.8515 ± 0.3435	0.0311 ± 0.0440
0.005	0.001	0.3258 ± 0.1247	0.2070 ± 0.1320	0.2753 ± 0.0901
0.005	0.008	0.5606 ± 0.1101	0.3815 ± 0.1988	0.0718 ± 0.0696
0.005	0.032	0.5680 ± 0.2099	0.4187 ± 0.2275	0.0013 ± 0.0018

NEC remains 0.03; other settings match Table 31. Responses use four evaluated windows and all 20 perturbation sources. δ affects both training and response evaluation. Within/between quantities describe true-state groups; floor frequency uses mean absolute response $\leq 10^{-6}$. Their magnitudes change with perturbation amplitude, so they are reported descriptively.

E.2 Fixed-support fitting and one-pass refinement

Table 33 asks how much error remains when support and state responsibilities are held fixed. On the development graph, the true support gives substantially lower validation MSE than either learned support under the same fresh decoder and budget.

For four additional graphs, I2 and I3 freeze the PCDM and Independent inferred responsibilities. Each target considers 145 supports, including the unchanged support, with identical fresh-decoder refits. The 3,071 training targets are split into 2,456 fitting and 615 selection targets. Candidate selection uses observation MSE; state labels and true edges are used only for subsequent diagnostics. Frozen responsibilities retain the provider's training context. Across two providers, four graphs and 20 targets, the search costs 23,200 candidate fits of 90 epochs/1,080 updates each.

Refinement lowers aggregate validation MSE and improves mean PRIVATE AP for both providers. I2 nevertheless adds two false parents on graph 71, while its total false-coordinate count falls; its all-zero PRIVATE count on graph 37 rises from 17 to 20. Every run worsens some state–target MSEs (6–12 of 60 inferred-state groups and 7–14 true-state groups). Binary-support scores $0.25H$ remain distinct from the continuous effective scores used by neural models.

Table 33: Fixed-support function fitting and one-pass support refinement.

(a) Given states, common decoder, one development graph

Support	Train MSE ↓	Val. MSE ↓	Mean error ↓
H_2	0.455320	0.462414	0.444693
H_3	0.438027	0.446021	0.427916
H_{true}	0.155436	0.160441	0.142521

(b) Frozen inferred states, four additional graphs

Start / refit	Val. MSE ↓	PRIVATE AP ↑	Owner (%) ↑	Correct/60 ↑	FP/180 ↓
I2 / unchanged	0.4699 ± 0.0457	0.6130 ± 0.0350	66.18 ± 3.47	31.2 ± 3.4	82.2 ± 11.5
I2 / refined	0.3918 ± 0.0452	0.7215 ± 0.0450	75.56 ± 3.51	39.5 ± 3.0	64.0 ± 12.5
I3 / unchanged	0.4666 ± 0.0342	0.6264 ± 0.0201	67.50 ± 2.12	32.8 ± 2.6	79.5 ± 8.8
I3 / refined	0.3861 ± 0.0371	0.7243 ± 0.0267	76.46 ± 2.88	40.5 ± 2.5	60.8 ± 9.6

Panel (a): same fresh gain-four, unconstrained decoder and true-state responsibilities, 90 epochs/1,080 updates, fixed endpoint; Mean error is validation conditional-mean MSE. H_2/H_3 come from PCDM/Independent. Panel (b): I2/I3 denote their respective inferred-state providers; each refinement is compared with its own equally refitted unchanged support. Scores are $0.25H$. Owner splits credit among tied state maxima; Correct requires a unique correct maximum. Panel (b) shows mean ± sample SD for graphs 23/37/53/71.

E.3 Auxiliary response consistency

CNC applies a common shift to one uniformly drawn lagged source per window: $\delta_b = s_b\lambda[Q_{0.95}(x_{b,:,p_b}) - Q_{0.05}(x_{b,:,p_b})]$, with random sign s_b. The factual posterior, including its gradient path, is reused for both decoder evaluations. For their difference $\Delta Y_{b,t}$, define

$$c_{b,t} = \max\left(d^{-1}\sum_i |\Delta Y_{b,t,i}|, \epsilon\right), \qquad \widetilde{\Delta Y}_{b,t} = \Delta Y_{b,t}/\operatorname{stopgrad}(c_{b,t}),$$

$$L_{\text{invariant}} = d^{-1}\mathbb{E}_{(t_1,t_2)}\|\widetilde{\Delta Y}_{b,t_1} - \widetilde{\Delta Y}_{b,t_2}\|_2^2.$$

Pairs are sampled within valid MAP states, falling back to valid times in the same sequence when insufficient. CNC-ON uses weight 0.001, perturbation fraction 0.008, two pairs, all outputs and floor 10^{-6} after five adaptation epochs. One backward pass adds decoder and posterior gradients before their separate clipping and Adam updates; residual-teacher responsibilities remain detached. CNC-FG differentiates the denominator, retaining Top-k/STE approximations.

The four-arm controlled comparison uses the five additive development tuples, unit-gain spectrally normalized VIS, matched initializations and the same 90-epoch budget: 1,080 decoder and 1,020 posterior updates. NEC and DAG penalties are zero. Neural-q validation MSE selects the earliest minimum, followed by one fixed validation permutation. The comparison evaluates validation responses/states and fixed-mechanism observations.

Observable-matched pairing. CNC-Matched preserves CNC-ON's window, source, sign, magnitude and absolute time gap, and draws uniformly from eligible pairs in the same parent-input RMS and physical-source-distance bins. Bin edges are $[0, 0.25, 0.5, 1, 2, 4, \infty)$ and $[0, 0.1, 0.25, 0.5, 1, 2, 4, \infty)$, with left-inclusive membership. A separate random stream is used; singleton pools remain and bins are never relaxed. States do not filter candidates. Coverage is 100%, the mean pool size is 4.99, and 65.59% of pairs differ from CNC-ON. True-state purity remains 90.94%, versus 92.98% for CNC-ON, so matching preserves considerable state similarity.

Response and state measurements. The bank covers 191 training and 48 validation windows, all 20 sources and both signs. A physical target appearing n_t times receives weight $1/n_t$ per exposure. Predicted and true finite responses use train-standardized units; the generator evaluates the native float32 perturbation, converted to physical units, in float64. Response NMSE retains all zero predictions and equals one for an all-zero predictor when true response energy is positive. Inactive-coordinate response energy is reported directly; normalized direction summaries report above-floor coverage. State probabilities are averaged over window exposures before argmax, and boundary matching maximizes one-to-one matches within two targets. PRIVATE AP and owner retain all 60 edges and exact ties.

Table 34: Validation outcomes: mean ± sample SD over five paired tuples. BA and owner are percentages.

Arm	Response NMSE ↓	ARI ↑	BA (%) ↑	PRIVATE AP ↑	Owner (%) ↑
CNC-OFF	0.9839±0.0174	0.656±0.262	80.53±16.23	0.508±0.051	48.78±7.44
CNC-ON	**0.9834±0.0167**	0.659±0.266	81.07±15.96	**0.514±0.046**	**52.33±3.34**
CNC-Matched	0.9841±0.0178	**0.666±0.257**	81.06±16.03	0.499±0.054	48.00±8.99
CNC-FG	0.9850±0.0187	0.659±0.266	**81.28±15.97**	0.503±0.045	48.78±6.45

CNC-ON uses inferred-state pairs; CNC-Matched uses observable-matched pairs. Both hold the denominator fixed in backpropagation (SG). CNC-FG differentiates it (FG).

CNC-ON minus CNC-OFF gives response NMSE -0.000516 $[-0.004242, 0.003211]$ and PRIVATE AP $+0.00591$ $[-0.01942, 0.03124]$. Balanced accuracy rises by 0.5348 $[0.0315, 1.0381]$ percentage points, but its contrast with CNC-Matched includes zero. Owner and boundary-F1 intervals also include zero. CNC-FG increases ABSENT selections relative to CNC-ON by 1.7143 $[0.2980, 3.1306]$ points and yields fewer uniquely localized private edges. These are exploratory uncorrected intervals over five tuples. The true-function bank contains 41–42% opposite-sign pairs within the same mechanism: common nonlinear mechanisms can produce different signed responses at different inputs.

Each arm reads three fixed-mechanism trajectories per graph with 6,143 targets and no true boundaries, keeping its trained parameters, scaler and permutation. CNC-OFF, ON, Matched and FG retain respectively 132.04, 128.71, 122.07 and 123.11 false boundaries per 1,000 targets. All remain oversegmented; constant-state ARI and boundary PRF are undefined.

Normalization check. The fixed-response check holds outputs, pair identities and state weights fixed for 15 selected checkpoints and 4,800 training pairs. SG and FG forward values agree exactly. Float64 radial residuals are at most 2.85×10^{-14} for SG and 1.16×10^{-14} for nonfloor FG; the ordinary FG finite-difference residual is at most 7.11×10^{-10}. Both-floor and mixed-floor pairs reach 5.00% and 2.19% and use their corresponding derivative branches. The theoretical identities concern fixed responses; training also changes states and pairs.

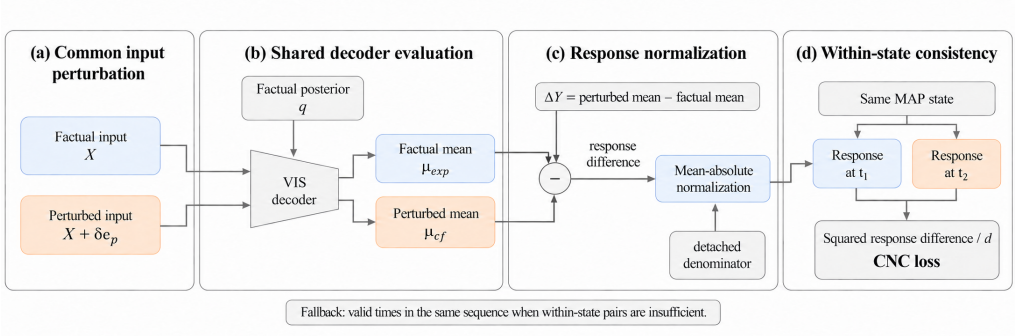

Figure 3: Auxiliary response consistency. CNC compares normalized decoder responses under a common input perturbation and the factual posterior. The normalization denominator is held fixed during backpropagation. This objective is used only in the supplementary comparisons.

E.4 Optional necessity constraint

NEC is an optional deletion-based ranking term for the auxiliary analyses in this appendix. It is off in the primary learner and in the controlled CNC comparison. Each batch samples a target uniformly and one of its top-m candidate parents. Deleting an edge from all states changes short-rollout prediction error by

$$\Delta_e^{\mathrm{rel}} = \frac{\mathcal{E}(A \setminus e) - \mathcal{E}(A)}{\mathcal{E}(A) + \epsilon}.$$

Standardizing these costs across sampled candidates and applying softmax with temperature τ_{nec} gives a target distribution p_e. The same transformation of continuous edge scores gives q_e. The loss is

$$L_{\mathrm{NEC}} = \frac{1}{E'} \sum_{e \in \mathcal{C}} p_e \log(p_e/q_e),$$

where E' counts retained draws, including duplicates. Deletion costs and the reused factual posterior are detached; gradients pass only through q_e. The controlled settings use weight 0.03, four draws, top-three candidates, horizon two and temperature 0.2.

E.5 Conditional response-error diagnostics

For true and predicted responses $R, \widehat{R}$ under inverse-overlap weights, let $T = \|R\|_w^2$, $P = \|\widehat{R}\|_w^2$, $C = \langle R, \widehat{R} \rangle_w$, $r = \sqrt{P/T}$ and $\gamma = C/\sqrt{PT}$. Then $\mathrm{NMSE} = 1 + r^2 - 2r\gamma$. An oracle nonnegative output multiplier leaves error $1 - \max(0, C)^2/(PT)$ when $P, T > 0$. If $P = 0 < T$, NMSE is one; if $T = 0$, relative errors are undefined. On CNC validation responses, RMS ratios are 0.362–0.365 and alignments 0.202–0.204; optimal scaling still leaves NMSE about 0.958.

With factual responsibilities fixed, a coordinate absent from every state's hard support has exactly zero predicted response. Its true-response energy gives the support lower bound used above. For true-state branch response R_z, the fit error $R_z - R$ and mixing error $\widehat{R} - R_z$ sum to the total response error; their weighted squared norms include a cross term, which is negative in these evaluations. These terms are algebraic components, rather than exclusive causal fractions. Raw gates span 0.239–0.753 while hard masks exclude 84.21% of inputs, locating the zero scores in support selection.

Conditional splines on the development tuples attain fresh response NMSE 0.6816 ± 0.0771 with the original support and frozen CNC-OFF posterior (S0), 0.1200 ± 0.0777 with all 19 non-self inputs and the same posterior (S1), and 0.0197 ± 0.0006 with all inputs and true states (S2). They use the spline basis and ridge selection defined above; S2 uses given validation states for selection. All 45 candidates are fitted (2,700 state–target systems), and selected candidates all use $\lambda = 0.001$ without another fit. No true support, weights or sinusoidal features enter the fits. These references change function family, capacity, support or state information and retain the unresolved neural fitting error.

Supplementary materials. The anonymous supplement includes executable implementations, bundled datasets and configurations. Separate commands run controlled-experiment training, technical checks and reaggregation of published results. The manuscript PDF, complete LaTeX source, figure files and a concise usage guide are included.