When Does Forecasting Reveal Temporal Structure? Recoverability and Stability of Forecast-only Explanations

Anonymous authors
Paper under double-blind review

Abstract

Forecast accuracy is often used as a proxy for temporal structure discovery, but predictive performance and structural recoverability are not equivalent. Forecast-only structural selection—choosing among candidate temporal structures based solely on their forecasting performance—provides a practical way to identify predictive temporal dependencies for model interpretation, diagnosis, and downstream decision making. In this work, we study when forecast-only structural selection can be trusted. We show that a vanishing forecast margin does not necessarily imply structural ambiguity, and establish a stability perspective that relates reliable forecast-only structural selection to the margin's robustness against uncertainty in the selection objective. Experiments across controlled and end-to-end settings demonstrate that forecast margin alone is insufficient, while the proposed stability condition provides a sufficient guarantee for correct forecast-only structural selection. We further give a model-based certification procedure that uses independent data to check the selected structure under a declared structural family, abstains when the evidence is insufficient, and does not require knowing the candidates' ideal risks. Our results suggest that predictive accuracy can support structural interpretation when its advantage is sufficiently robust, or when the selected structure is independently certified under a declared structural family. More broadly, our framework provides a principled basis for evaluating when forecast-derived temporal structures can be trusted as model explanations in theoretical analysis and finite-candidate time-series applications.

1 Introduction

Modern time-series models are increasingly expected not only to predict accurately, but also to reveal which parts of the observed history support their predictions. Recent approaches therefore use temporal masks, perturbations, attention, and information bottlenecks to identify influential time points or dependencies (Ozyegen et al., 2022; Liu et al., 2024b;a; Zheng et al., 2026). Such forecast-derived temporal structures are often used for interpretation, diagnosis, or downstream decision making. This raises a basic question: does accurate forecasting imply that the reported temporal structure is correct?

In general, predictive performance need not uniquely determine the mechanism or explanation used by a model, as emphasized by work on underspecification, predictive multiplicity, and explanation faithfulness (D'Amour et al., 2022; Marx et al., 2020; Jain & Wallace, 2019; Jacovi & Goldberg, 2020). The issue is particularly acute in time series, where nearby lags can be strongly correlated and different parts of the history can provide redundant predictive information. Recent methods assess which temporal information a predictor uses or needs, for example through learned masks or forecast-necessity tests (Zheng et al., 2026; Kuskova et al., 2026a). What remains less understood is how differences in forecast performance between candidate temporal structures (*forecast margin*) relate to the ability to recover the true structure from the full observed trajectory (*structural recoverability*), and when forecast-only structural selection can be trusted.

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.

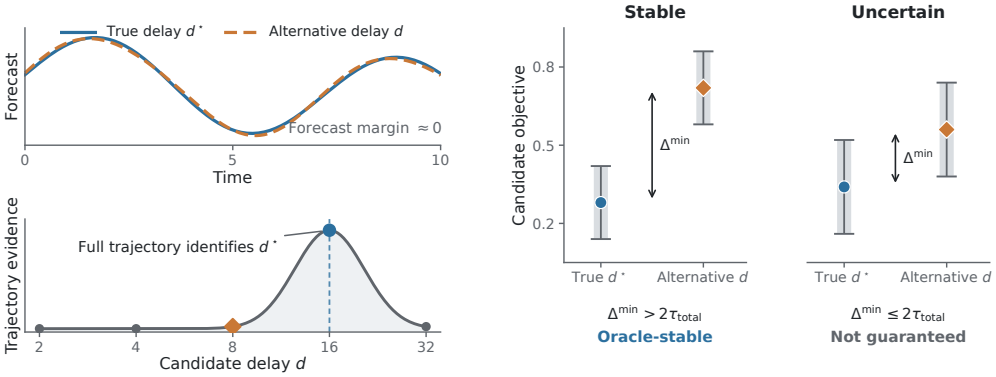
We first study this question through temporal delay, a basic and widely used form of temporal structure, while the stability formulation developed later applies more generally to finite candidate structural families.

We separate the problem into three questions:

$$\boxed{\text{recoverability} \longrightarrow \text{forecast margin} \longrightarrow \text{structural stability}}.$$

Recoverability asks whether the full observed trajectory contains enough information to distinguish the true temporal structure from the candidate alternatives. The **forecast margin** is the increase in the best achievable per-observation forecast MSE when an alternative structure is used. The **structural stability** question concerns whether the objective used to choose candidate structures preserves the true structure as its top-ranked candidate after training and finite-sample evaluation.

Forecast similarity and structural recoverability are fundamentally different notions. As illustrated in Figure 1(a), two candidate temporal structures can achieve nearly identical forecast performance even when the full trajectory still clearly identifies the true delay. Thus, a small forecast margin does not necessarily imply structural ambiguity.



(a) Forecast similarity vs. structural recoverability.

(b) Forecast margin vs. selection stability.

Figure 1: **Forecast accuracy alone does not determine whether a temporal structure can be trusted. (a)** Two candidate delays can have nearly identical forecast performance, while the full observed trajectory still clearly identifies the true delay $d^\star$. **(b)** Structural selection is oracle-stable when the forecast margin is large enough to remain reliable despite variation from fitting and evaluation; otherwise, the oracle condition gives no guarantee.

This distinction motivates a stability perspective on forecast-only structural selection. As illustrated in Figure 1(b), when temporal structure is selected by forecast performance, the relevant question is whether the predictive advantage of the true structure is large enough to survive uncertainty in the empirical selection objective. We formalize this through a stability ratio comparing forecast margin with objective distortion, which yields a sufficient condition for correct structural selection.

Our contributions are summarized as follows:

- We characterize the **relationship between structural recoverability and forecast margins** in controlled Gaussian temporal models, showing that vanishing per-observation forecast differences can coexist with reliable recovery.
- We establish a **sufficient stability guarantee for forecast-only structural selection**, connecting ideal forecast margins to distortion in the empirical selection objective. Controlled and learned-predictor experiments examine its scope and conservativeness.
- We provide a **computable post-selection structural certificate under explicit model assumptions**. The procedure treats the forecast-selected structure as a proposal, applies conditional split/profile likelihood inference (Wasserman et al., 2020) to independent data, and either certifies the proposal or abstains.

2 Related Work

Time-delay estimation and recovery. Delay and lead–lag estimation are classical problems in system identification and time-series analysis, where recovery accuracy depends on signal strength, noise, and temporal dependence among candidate lags (Bjorklund & Ljung, 2003; Sadler & Kozick, 2006; Ljung, 1999; Ma & Huang, 2023; Cattaldo et al., 2024; Heyse et al., 2021; Zhao & Shen, 2024; Runge et al., 2019a;b; Runge, 2020). Kuskova et al. (2026b) give a pre-fit diagnostic for interaction recoverability under dependent lagged inputs. We connect this recovery question to forecast-only structural selection through trajectory information, candidate forecast margins, and uncertainty in the selection objective.

Learning temporal dependence for forecasting. Time-series models encode temporal dependence through masks, selected context, attention, information bottlenecks, and multiscale spatiotemporal representations (Ozyegen et al., 2022; Liu et al., 2024b;a; Zhang et al., 2025; Qian & Mallick, 2024). IB-Forecast learns explicit input masks (Zheng et al., 2026), while DSPR learns regime-dependent history together with interaction structure (Zhang et al., 2026). Forecast-necessity testing removes candidate temporal relations and checks how much forecast accuracy changes (Kuskova et al., 2026a). These methods ask which history a predictor uses or needs. We instead compare a prespecified family of candidate structures and ask when their forecast performance supports correct structural selection.

Prediction accuracy and model structure. Different models can predict similarly while using different variables or decision rules, as studied in underspecification, predictive multiplicity, shortcut learning, and explanation faithfulness (D'Amour et al., 2022; Marx et al., 2020; Geirhos et al., 2020; Jain & Wallace, 2019; Jacovi & Goldberg, 2020; Ross et al., 2017). Nearby lags and past outputs can replace part of the same predictive signal (Tunyi, 2026). Statistical model selection distinguishes identification from regression estimation (Yang, 2005), and cross-validation consistency depends on sample allocation and the procedures compared (Yang, 2007). Best-subset-selection theory relates recovery to predictive separation after fitting alternatives under dependent designs (Guo et al., 2020; Roy et al., 2025). Our analysis connects trajectory information, forecast margins, and stable forecast-only structural selection. The post-selection certificate applies universal inference through split/profile and conditional likelihoods (Wasserman et al., 2020) to decide whether a forecast-selected structure can be certified under a declared structural family.

3 Problem Setup

We use two controlled mechanisms:

$$\text{P1:} \quad Y_t = \alpha U_{t-d} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2); \tag{1}$$

$$\text{P2:} \quad Y_t = aY_{t-1} + bU_{t-d} + \epsilon_t. \tag{2}$$

P1 isolates a single input delay. P2 adds output memory and lets an alternative delay re-fit the other coefficients. This makes P2 a richer setting for predictive substitution, since an alternative delay can exploit both correlated lag information and output memory after re-fitting.

Let L be the largest candidate lag and define

$$\mathcal{I}_L = \{L+1, \ldots, T\}, \qquad n_{\text{eff}} = T - L.$$

For a candidate delay d,

$$S_d u = (u_{t-d})_{t \in \mathcal{I}_L}$$

is the shifted input trajectory.

For P2 we use the shared window

$$\mathcal{I}_L^{\text{P2}} = \{L+2, \ldots, T\}, \qquad n_{\text{P2}} = T - L - 1, \tag{3}$$

so that every candidate uses the same observed history. Here n_{eff} and n_{P2} count the scored observations; they are not effective sample sizes adjusted for temporal dependence. The exact assumptions are listed in Appendix A.

4 THEORY: RECOVERABILITY, FORECAST MARGIN, AND STABILITY

P1 relates trajectory-level recovery to the per-observation forecast margin. P2 extends this relation to fitted candidate structures, where alternative delays may adapt through re-fitting. The oracle stability condition then asks when the empirical forecast-selection objective preserves the ideal structural ordering. Finally, the computable certification step treats the forecast-selected structure as a proposal and asks whether independent data can certify it under a declared structural family.

4.1 P1: PREDICTION CAN BE ACCURATE WHILE STRUCTURE REMAINS RECOVERABLE

Consider two equally likely delays d_0, d_1 under P1, conditional on a fixed input trajectory u. Their output laws have means $\mu_d = \alpha S_d u$ and common covariance $\sigma^2 I$. The optimal recovery rule selects the larger likelihood, equivalently the smaller empirical MSE. Define the per-observation excess forecast risk

$$\Delta_{n_{\text{eff}}}(d_1; d_0, u) := r_{n_{\text{eff}}}(d_1) - r_{n_{\text{eff}}}(d_0) = \frac{\alpha^2}{n_{\text{eff}}} \|S_{d_0} u - S_{d_1} u\|_2^2. \tag{4}$$

This is the true candidate's forecast advantage, not its total MSE.

Proposition 4.1 (P1: recoverable structure with vanishing forecast advantage)**.** *Under P1 with independent Gaussian noise of variance σ^2 and equal priors,*

$$D_{\text{KL}}(d_0 \| d_1 \mid u) = \frac{n_{\text{eff}} \Delta_{n_{\text{eff}}}(d_1; d_0, u)}{2\sigma^2}, \tag{5}$$

$$e^\star_{\text{rec}}(d_0, d_1; u) = \Phi\left(-\frac{\sqrt{n_{\text{eff}} \Delta_{n_{\text{eff}}}(d_1; d_0, u)}}{2\sigma} \right), \tag{6}$$

$$r_{n_{\text{eff}}}(d_0) = \sigma^2, \qquad r_{n_{\text{eff}}}(d_1) = \sigma^2 + \Delta_{n_{\text{eff}}}(d_1; d_0, u), \tag{7}$$

where Φ is the standard normal cumulative distribution function. In particular, along a sequence of problems,

$$\Delta_{n_{\text{eff}}} \longrightarrow 0, \qquad n_{\text{eff}} \Delta_{n_{\text{eff}}} \longrightarrow \infty \quad \Longrightarrow \quad e^\star_{\text{rec}} \longrightarrow 0.$$

Thus the excess forecast risk vanishes while the true delay remains consistently recoverable. The proof is in Appendix B.5.

The proposition separates the size of a predictive advantage from the evidence available for a structural decision. Recovery depends on accumulated evidence $n_{\text{eff}} \Delta_{n_{\text{eff}}}$, whereas the excess forecast risk is normalized per observation. Consequently, a wrong delay can become arbitrarily close in population MSE without becoming structurally equivalent to the true delay.

This is not a separation between likelihood selection and empirical-MSE selection: they coincide in this setting. It is a warning against interpreting the population margin alone as statistical ambiguity. The limiting statement concerns a local sequence of problems whose parameters may vary with n; collecting more observations from one fixed system need not shrink its population margin (Appendix A).

4.2 P2: THE FORECAST MARGIN AFTER THE WRONG DELAY IS RE-FITTED

P1 compares fixed input coefficients, whereas in P2 each candidate delay is allowed to re-fit both the output-memory coefficient and the input gain. The wrong delay can then use past output and a correlated input lag as predictive substitutes. The relevant margin is what remains after this adaptation, not the loss incurred by changing the delay while leaving the coefficients fixed.

With risk $\mathcal{R}$ averaged over the common scoring window under the true law (Appendix A), define

$$r^\star(d) := \inf_{a,b} \mathcal{R}(d, a, b), \qquad \Delta_{\text{prof},\text{n}_{\text{P2}}}(d) := r^\star(d) - r^\star(d^\star).$$

Let $P^\star$ denote the true trajectory law and P_{d,a,b,s^2} the candidate conditional Gaussian law. We distinguish candidates constrained to use the true innovation variance from those allowed to fit it:

$$J^{\text{fixed}}_{\text{prof}}(d^\star, d) := \inf_{a,b} D_{\text{KL}}(P^\star \| P_{d,a,b,\sigma^2}),$$

$$J^{\text{free}}_{\text{prof}}(d^\star, d) := \inf_{a,b,s^2} D_{\text{KL}}(P^\star \| P_{d,a,b,s^2}).$$

Proposition 4.2 (P2: information after candidate re-fitting)**.** *Under the common boundary/input law in Appendix A,*

$$J_{\mathrm{prof}}^{\mathrm{fixed}}(d^\star, d) = \frac{n_{\mathrm{P2}}\Delta_{\mathrm{prof},\mathrm{n_{P2}}}(d)}{2\sigma^2}, \tag{8}$$

$$J_{\mathrm{prof}}^{\mathrm{free}}(d^\star, d) = \frac{n_{\mathrm{P2}}}{2}\log\left(1 + \frac{\Delta_{\mathrm{prof},\mathrm{n_{P2}}}(d)}{\sigma^2}\right). \tag{9}$$

The proof is in Appendix B.6.

Re-fitting cannot increase wrong-delay risk relative to fixing its coefficients in the same P2 system. To see what signal survives this adaptation, consider centered stationary Gaussian P2 with nonsingular predictor covariance. The projection calculation in Appendix B.7 gives

$$\Delta_{\mathrm{prof}}(d) = (b^\star)^2 \operatorname{Var}(U_{t-d^\star} \mid Y_{t-1}, U_{t-d}). \tag{10}$$

The conditional variance measures the part of the true delayed input that neither the wrong lag nor past output can reproduce. Predictive substitution reduces this residual, rather than making the two delay labels identical. Its partial-correlation factorization and the numerical projection check are given in Appendices B.7 and E.2.

Proposition 4.2 connects the remaining residual risk to accumulated trajectory separation. Unlike the fully specified binary P1 comparison, this profiled KL does not itself give an exact composite-hypothesis recovery error; that distinction is maintained in the experiments.

4.3 From the ideal forecast margin to the objective used for selection

Finite training and selection replace the ideal risk $r^\star(d)$ by an empirical objective $\widehat{r}(d)$. These are two distinct changes. If R_d is the population risk of the frozen fitted predictor, then

$$\widehat{r}(d) - r^\star(d) = \underbrace{R_d - r^\star(d)}_{\text{fitted-predictor gap}} + \underbrace{\widehat{r}(d) - R_d}_{\text{finite-selection fluctuation}}. \tag{11}$$

The first term includes the predictor class, finite training data, and the fitting procedure; it is not solely optimization error. The second changes when the same frozen predictors are evaluated on a new finite trajectory. A fitted wrong-delay predictor can therefore have lower population risk than a poorly fitted true-delay predictor, even when the ideal candidate ordering is correct. Additional selection observations change the evaluation fluctuation, not those frozen population risks. We ask when the combined effect preserves the true candidate as the objective's minimizer.

Define the *total objective distortion*

$$\tau_{\mathrm{total}} := \max_d \left|\widehat{r}(d) - r^\star(d)\right|. \tag{12}$$

Let the minimum ideal forecast margin be

$$\Delta^{\min} := \min_{d \neq d^\star} \left[r^\star(d) - r^\star(d^\star)\right] > 0. \tag{13}$$

For P2, this is the minimum profiled forecast margin.

The following theorem guarantees that the true candidate remains the unique minimizer.

Theorem 4.3 (A margin larger than twice the total objective distortion preserves the true structure)**.** *If*

$$\Delta^{\min} > 2\tau_{\mathrm{total}}, \tag{14}$$

then $d^\star$ is the unique minimizer of $\widehat{r}$. The factor 2 is sharp in the worst case: if $\Delta^{\min} < 2\tau_{\mathrm{total}}$, an allowed distortion can make an alternative structure have a smaller objective than the true one; at equality it can create a tie. The proof is shown in Appendix B.9.

An ε-optimal ideal-risk guarantee ensures correct structure when $\varepsilon < \Delta^{\min}$; for $\varepsilon \geq \Delta^{\min}$ that guarantee alone is inconclusive (Proposition C.5).

We then define *stability ratio*

$$S_{\mathrm{total}} := \frac{\Delta^{\min}}{2\tau_{\mathrm{total}}}. \tag{15}$$

Set $S_{\text{total}} = +\infty$ when $\tau_{\text{total}} = 0$. Then $S_{\text{total}} > 1$ guarantees correct selection, whereas $S_{\text{total}} \leq 1$ is inconclusive. The bound controls absolute objective distortions, although selection depends on differences: a shared shift can change τ_{total} without changing the selected candidate. Thus this is an oracle sufficient condition, not a universal difficulty scale. Section 5.3 examines its conservativeness; Appendix G.4 gives an explicit opposite-trend example.

4.4 Computable structural certification

The oracle condition explains when forecast-only selection is safe in principle, but it cannot be checked without ideal risks. We therefore treat the forecast-selected structure as a proposal and apply split/profile likelihood inference (Wasserman et al., 2020) to an independent certification trajectory. Training data A fit the predictors; selection data B choose $\widehat{d}$ by forecast MSE. Before observing certification outcomes C, freeze $\widehat{d}$, its predictor, and a positive variance estimate. They define a normalized conditional trajectory density $Q(C)$.

For candidate d, let $\mathcal{H}_d$ contain Gaussian transitions with mean $c + aY_{t-1} + bU_{t-d}$ and unknown variance $s^2 > 0$. Conditional likelihoods share the observed exogenous input and initial history. Define

$$E_d = \frac{Q(C)}{\sup_{\theta \in \mathcal{H}_d} p_{d,\theta}(C)}, \qquad \mathcal{C}_\alpha = \{d : E_d < 1/\alpha\}. \tag{16}$$

The numerator evaluates the already fitted predictor; the denominator gives candidate d its best possible fit within the entire declared structural family. Thus a candidate is challenged even after its coefficients and variance can adapt to the certification data. Least squares and variance profiling compute this denominator without the true structure, true noise variance, or ideal risks.

Certify the original answer only if $\mathcal{C}_\alpha = \{\widehat{d}\}$. Multiple retained candidates indicate unresolved alternatives; an empty set or a singleton disagreeing with $\widehat{d}$ also leads to abstention. This rule does not replace or correct the original forecast-only answer.

Proposition 4.4 (Conditional split-likelihood certification)**.** *Suppose the true conditional law belongs to $\mathcal{H}_{d^\star}$, the certification trajectory is independent of A, B, and Q is a normalized density frozen before its outcomes. Under the shared boundary and exogeneity conditions,*

$$\Pr(d^\star \notin \mathcal{C}_\alpha) \leq \alpha, \qquad \Pr(\widehat{d} \neq d^\star,\ \mathcal{C}_\alpha = \{\widehat{d}\}) \leq \alpha. \tag{17}$$

The key inequality is $E_{d^\star} \leq Q(C)/p_{d^\star,\theta^\star}(C)$. Conditional on pre-certification information and the common input/boundary variables, the right-hand side has expectation one; Markov's inequality gives the result. This is a trajectory-density argument, not an assumption that all scored transitions are independent.

The bound controls erroneous certification jointly, not error conditional on certification. An MLP may supply Q, but the verified structural families remain linear Gaussian; a local neural fit cannot replace the denominator supremum. This application of universal inference is a post-selection, model-based certificate for the declared structural family, not a distribution-free estimate of S_{total}. Proof, closed-form calculation, and scope are in Appendix H.

5 Experiments

5.1 Forecast-only structural selection protocol

For each candidate delay d, fit a separate predictor $\widehat{f}_d$ using only the inputs permitted by that candidate. All fitted candidates are evaluated on the same selection trajectory, independent of training:

$$\widehat{r}(d) = \frac{1}{n_{\text{sel}}} \sum_{i=1}^{n_{\text{sel}}} (Y_i - \widehat{f}_d(X_i))^2, \qquad \widehat{d} = \arg\min_d \widehat{r}(d). \tag{18}$$

Within a continuous trajectory, observations may be temporally dependent. OLS provides a well-specified P2 baseline; the candidate-restricted MLP uses the same inputs and selection rule, separating the choice of learner from the structural comparison.

The P1 known-mean experiment evaluates the information–margin identity directly. Learned-candidate experiments distinguish ideal risks, fitted-predictor population risks, and finite selection. True delays and ideal risks are used only for evaluation, not training or selection. The computable certification experiment adds an independent trajectory and never accesses those oracle quantities in its decision. Full protocols and the semi-synthetic process illustration are in Appendices E, I, and F.

5.2 P1: RECOVERY REMAINS POSSIBLE AS THE FORECAST-MSE GAP VANISHES

This experiment tests whether a vanishing forecast margin must imply structural ambiguity. We evaluate known-mean P1 candidates along the local sequence

$$\alpha_n = c\, n^{-\beta/2}, \qquad n = n_{\mathrm{eff}}, \qquad \beta = 0.5,$$

for which the per-observation margin vanishes while accumulated information grows. At each sample size, the analytic Bayes error is evaluated conditional on the observed input and compared with noisy-trajectory MAP decisions. The protocol uses independent input trajectories and repeated decisions, as specified in Appendix E; there is no fitted neural predictor in this comparison.

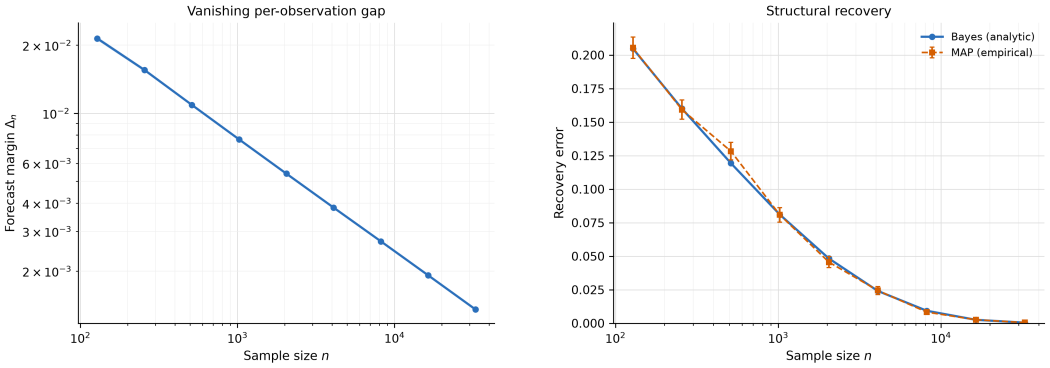


Figure 2: P1 experiment: forecast margin vanishes while structural recovery remains possible. (a) The forecast margin Δ_n decreases with sample size. (b) Structural recovery error decreases, consistent with the predicted limit of 0.

Figure 2(a) shows decreasing excess forecast risk, whereas Figure 2(b) shows decreasing structural recovery error. Their simultaneous decline is the relevant observation: the alternative becomes closer in average prediction quality even as the trajectory provides stronger evidence for the true delay. Agreement between the analytic and empirical error curves checks the exact relation in Proposition 4.1.

The interpretation depends on the local-sequence design. Signal strength changes together with sample size; this is not evidence that collecting more data from one fixed mechanism shrinks its population margin. Moreover, MAP and empirical-MSE selection coincide for these fixed Gaussian candidates. The result separates per-observation predictive similarity from statistical ambiguity, not two competing selection algorithms.

5.3 SCOPE AND CONSERVATIVENESS OF THE STABILITY GUARANTEE

Controlled candidate objectives. We first separate objective perturbation from model fitting using controlled P1 losses whose ideal risks are known. The study varies forecast margin, objective sample size, and noise over 90 conditions with 5,000 independent replicates each. Figure 3 groups the resulting selections by S_{total}. All 97,238 samples above the sufficient boundary select correctly; below it, correct and incorrect choices coexist. Complete counts and generation details are in Appendix E.4.

The downward trend in Figure 3 describes this grid, not a universal ordering of difficulty. The zero-error region checks the deterministic implication of Theorem 4.3; it is not independent evidence that a new selector is more accurate. Shared objective fluctuations can make the ratio conservative, and Appendix G.4 exhibits improving recovery with a decreasing typical ratio.

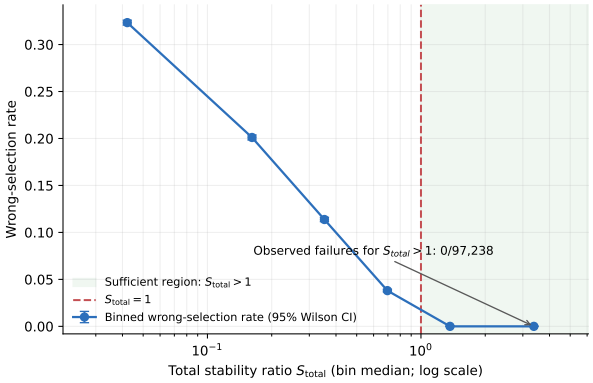


Figure 3: P1 wrong-selection rate by S_{total} bin, with 95% Wilson intervals. No errors occur above 1; the binned trend is grid-specific.

Table 1: Oracle sufficient-region frequency on continuous P2 trajectories. Counts share fitted pairs and prefixes.

	OLS	MLP
Selections	864	864
$S_{\text{total}} > 1$	36.92%	28.47%
Errors in region	0	0

Learned candidates on continuous P2 trajectories. We then allow fitting and evaluation to contribute jointly to objective distortion. The experiment uses 12 stationary P2 mechanisms, training lengths $64, 256, 1024$, and selection lengths $32, 128, 512$, with eight independent blocks per mechanism. Each block has independent training and selection trajectories; lengths share prefixes, and candidates are evaluated on common observations. Table 1 reports the oracle sufficient-region frequency, not the coverage of the computable confidence set introduced later. The fitted-model and selection counts are therefore distinct statistical units (Appendix E.6).

Table 1 shows that the sufficient condition covers only a subset of selections. For OLS, 457 of the 545 choices outside that region are nevertheless correct. To distinguish conservativeness of the oracle sufficient condition from actual fitting and evaluation failures, Table 2 compares the population ordering of each frozen predictor pair with its finite-selection outcomes. Population risks are computed analytically for OLS and audited independently for MLP; unresolved Monte Carlo orderings are kept separate.

Table 2: Frozen-predictor population ordering versus finite selection on continuous P2 trajectories. Selection sizes reuse fitted pairs; the three unresolved MLP outcomes share one pair.

Population ordering	Empirical selection	OLS	MLP
Correct	Correct	761	734
Correct	Wrong	67	79
Wrong	Wrong	21	29
Wrong	Correct	15	19
Unresolved or tied	Either	0	3
Total		864	864

The off-diagonal outcomes in Table 2 expose the two effects in equation 11: a correct frozen-predictor ordering can be reversed by finite evaluation, and an already inverted population ordering can occasionally yield a correct finite-sample choice. The latter is not evidence that less evaluation data is preferable. Paired data-size results in Appendix E.8 further separate changes to fitted predictors from changes to their evaluation, without claiming a universal monotone improvement for every mechanism.

5.4 Post-selection certification without oracle risks

We test the post-selection certificate from Section 4.4 on five candidate delays using independent continuous P2 training, selection, and certification trajectories. Figure 4(a) varies certification length with the predictors and forecast-selected answers held fixed. The OLS and MLP mechanism grids

differ, so the curves are not a matched learner comparison. Figure 4(b) uses deliberately unequal predictor quality to distinguish predictive preference from structural certification. Protocols and counts are in Appendix I.

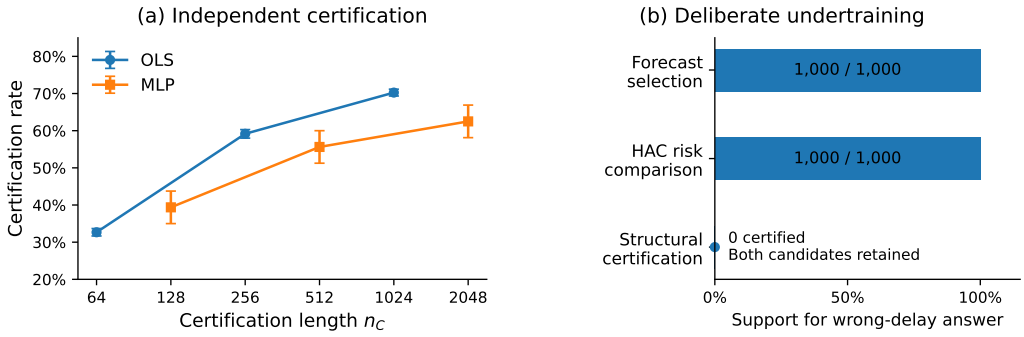


Figure 4: **Computable structural certification.** (a) Certification rate at $n_A = 256$, $n_B = 128$, and $\alpha = 0.05$; 95% intervals resample complete blocks within mechanisms. Lengths share prefixes, and the learner grids differ. (b) Deliberate undertraining separates predictive preference from structure: the structural procedure retains both candidates instead of certifying the wrong answer. Abstention is not correction.

Figure 4(a) shows that additional certification data supports more original choices in this grid: endpoint certification rates are 70.25% for OLS and 62.50% for MLP, with no observed erroneous certifications. High-correlation, high-noise conditions can remain inconclusive (Table 7 in the appendix). Figure 4(b) shows that predictive comparison can correctly prefer a lower-risk wrong-delay predictor while the structural check retains both candidates.

6 Discussion and Conclusion

Our results show that good forecasting does not necessarily mean that the correct temporal structure has been identified. In our controlled analysis, the forecast difference between two candidate structures can become very small even when the full trajectory still contains enough information to distinguish the true structure. Conversely, recoverable structural information does not guarantee that a finite-sample forecasting procedure will select the correct candidate. Therefore, when temporal structure is inferred from forecast performance, the key question is not only which structure predicts better, but whether that advantage is stable enough to be trusted.

This motivates evaluating forecast-only structural selection through stability rather than forecast margin alone. The condition $\Delta^{\min} > 2\tau_{\text{total}}$ gives a sufficient oracle-stability condition: when the ideal forecast margin dominates distortion in the empirical selection objective, the true structure remains the unique selected candidate. Importantly, failure to satisfy this condition does not imply that the selected structure is incorrect; it means that this sufficient condition alone does not guarantee correctness. Across controlled and end-to-end experiments, we examine the scope and conservativeness of this sufficient guarantee. The computable confidence set then provides a post-selection certificate for the forecast-selected delay under explicit model assumptions, using independent data rather than ideal risks.

The broader message is that prediction and explanation require different standards of evidence. A temporal structure should not be treated as an explanation simply because it achieves the lowest forecast error; its advantage over competing structures must also be stable enough to support that interpretation. Our results therefore shift the focus from asking only which structure predicts best to asking how strongly that structure is supported relative to plausible alternatives. Forecast-only explanations should report not only the selected structure, but also the strength, stability, and, when a declared structural family is available, certification status of the evidence behind that selection.

AI USE STATEMENT

Large language models were used as general-purpose research assistance for language polishing, identifying potentially relevant related work, and generating portions of the experimental code. References surfaced with AI assistance were checked by the authors against the original sources before citation. AI-generated code used in the reported experiments was manually reviewed, tested, and checked against the declared experimental protocol before execution. Reported numerical results were obtained by executing the reviewed code rather than by asking a language model to generate result values. The authors take full responsibility for the paper's claims, citations, code, and experimental results.

REPRODUCIBILITY STATEMENT

The appendix states the assumptions, exact risk and information formulas, proof details, data-generation rules, and experimental protocols. The reported experiments use fixed grids and fixed random seeds, with separate checks for population calculations, sampling, and structural selection. Structural selection uses forecast MSE only. The release records resolved configurations, seed maps, model hashes, candidate-wise selection losses, analytic margins, total objective distortion, and stability ratios. The independent certification study additionally records candidate-wise log likelihood ratios, retained sets, frozen variance estimates, independent-block identifiers, and the distinction between certification rate and erroneous-certification probability.

REFERENCES

S. Bjorklund and L. Ljung. A review of time-delay estimation techniques. In *42nd IEEE International Conference on Decision and Control (IEEE Cat. No.03CH37475)*, volume 3, pp. 2502–2507 Vol.3, 2003. doi: 10.1109/CDC.2003.1272997. URL https://doi.org/10.1109/CDC.2003.1272997.

Marco Cattaldo, Alberto Ferrer, and Ingrid Måge. Variable time delay estimation in continuous industrial processes. *Chemometrics and Intelligent Laboratory Systems*, 246:105082, 2024. doi: 10.1016/j.chemolab.2024.105082. URL https://doi.org/10.1016/j.chemolab.2024.105082.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1335.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Yongyi Guo, Ziwei Zhu, and Jianqing Fan. Best subset selection is robust against design dependence, 2020. URL https://arxiv.org/abs/2007.01478.

Jolan Heyse, Laurent Sheybani, Serge Vulliémoz, and Pieter van Mierlo. Evaluation of directed causality measures and lag estimations in multivariate time-series. *Frontiers in Systems Neuroscience*, 15:620338, 2021. doi: 10.3389/fnsys.2021.620338. URL https://doi.org/10.3389/fnsys.2021.620338.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://doi.org/10.18653/v1/N19-1357.

Valentina Kuskova, Dmitry Zaytsev, and Michael Coppedge. Beyond coefficients: Forecast-necessity testing for interpretable causal discovery in nonlinear time-series models. *The International FLAIRS Conference Proceedings*, 39(1), May 2026a. ISSN 2334-0754. doi: 10.32473/flairs.39.1.

Valentina Kuskova, Dmitry Zaytsev, and Michael Coppedge. When are neural interaction discoveries real? identifiability, recoverability, and a pre-fit diagnostic, 2026b. URL https://arxiv.org/abs/2606.08390.

Zichuan Liu, Tianchun Wang, Jimeng Shi, Xu Zheng, Zhuomin Chen, Lei Song, Wenqian Dong, Jayantha Obeysekera, Farhad Shirani, and Dongsheng Luo. TimeX++: Learning time-series explanations with information bottleneck. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32062–32082. PMLR, 2024a. URL https://proceedings.mlr.press/v235/liu24bl.html.

Zichuan Liu, Yingying Zhang, Tianchun Wang, Zefan Wang, Dongsheng Luo, Mengnan Du, Min Wu, Yi Wang, Chunlin Chen, Lunting Fan, and Qingsong Wen. Explaining time series via contrastive and locally sparse perturbations. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=qDdSRaOiyb.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, 2 edition, 1999. ISBN 9780136566953.

Xin-Yue Ma and Chun-Qing Huang. Data-driven approach for time-delay estimation of industrial processes. *ISA Transactions*, 137:35–58, 2023. doi: 10.1016/j.isatra.2023.01.028. URL https://doi.org/10.1016/j.isatra.2023.01.028.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Ozan Ozyegen, Igor Ilic, and Mucahit Cevik. Evaluation of interpretability methods for multivariate time series forecasting. *Applied Intelligence*, 52(5):4727–4743, 2022. doi: 10.1007/s10489-021-02662-2. URL https://doi.org/10.1007/s10489-021-02662-2.

Qipeng Qian and Tanwi Mallick. Wavelet-inspired multiscale graph convolutional recurrent network for traffic forecasting. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5680–5684, 2024. doi: 10.1109/ICASSP48485.2024.10446847.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations, 2017. URL https://arxiv.org/abs/1703.03717.

Saptarshi Roy, Ambuj Tewari, and Ziwei Zhu. Understanding best subset selection: A tale of two C(omplex)ities. *Electronic Journal of Statistics*, 19(1):2320–2342, 2025. doi: 10.1214/25-EJS2388. URL https://doi.org/10.1214/25-EJS2388.

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1388–1397. PMLR, 2020. URL https://proceedings.mlr.press/v124/runge20a.html.

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

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

Brian M. Sadler and Richard J. Kozick. A survey of time delay estimation performance bounds. In *2006 IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings*, pp. 282–288, Waltham, MA, USA, 2006. IEEE. doi: 10.1109/SAM.2006.1706138. URL https://doi.org/10.1109/SAM.2006.1706138.

Amadeo Tunyi. The failures of marginal influence-based attribution methods for global time series explanations. *arXiv preprint arXiv:2607.16236*, 2026. doi: 10.48550/arXiv.2607.16236. URL https://arxiv.org/abs/2607.16236.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117. URL https://doi.org/10.1073/pnas.1922664117.

Yuhong Yang. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika*, 92(4):937–950, 2005. doi: 10.1093/biomet/92.4.937. URL https://academic.oup.com/biomet/article/92/4/937/389439.

Yuhong Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007. doi: 10.1214/009053607000000514. URL https://arxiv.org/abs/0803.2963.

Jiahui Zhang, Zhengyang Zhou, Wenjie Du, and Yang Wang. Enhancing the maximum effective window for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://openreview.net/forum?id=Gmwsy7TlFI.

Yeran Zhang, Pengwei Yang, Guoqing Wang, and Tianyu Li. Dspr: Dual-stream physics-residual networks for trustworthy industrial time series forecasting, 2026. URL https://arxiv.org/abs/2604.07393.

Lifan Zhao and Yanyan Shen. Rethinking channel dependence for multivariate time series forecasting: Learning from leading indicators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b52b07a239a7afa155ca25cf17a55074-Abstract-Conference.html.

Xu Zheng, Wei Cheng, Zhuomin Chen, Mo Sha, Jingchao Ni, and Dongsheng Luo. Information bottleneck learning for faithful time series forecasting explanations. *arXiv preprint arXiv:2607.28124*, 2026. doi: 10.48550/arXiv.2607.28124. URL https://arxiv.org/abs/2607.28124.

Appendix organization. Appendix A states the contracts used by the theory. Appendix B contains the proofs, while Appendix C collects auxiliary statements used in those proofs. Appendix D gives the Bayes-oracle computation. Appendix E gives the experimental protocols for the oracle-stability analysis. Appendix F provides the semi-synthetic process illustration, and Appendix G gives auxiliary empirical checks. Appendices H and I provide the construction, guarantee, limitations, and full audit for the computable structural certificate.

A Detailed Assumptions and Notation

Candidate and input contract. The candidate delay is defined by the data-generating process. In the controlled P1 experiments, the input law does not depend on the candidate delay. Exact Bayes statements condition on every observed boundary variable used in the likelihood.

Noise contract. P1 uses independent Gaussian noise. P2 assumes ϵ_t is independent of the past output and observed exogenous input. For the exact P2 KL identities, the input trajectory and initial observed history have the same marginal law under every candidate; only the conditional transitions are changed. Candidate-specific stationary initial distributions would introduce an additional boundary KL term.

P2 boundary window. P2 uses Equation (3). This removes the first transition whose required past output is not included in the shared observed history.

Profiled P2 risk. For each candidate delay, the other linear coefficients are re-fitted before the candidate risks are compared. The resulting difference is the profiled forecast margin. Profile-KL is used as a population KL-divergence quantity; it is not called an exact composite-hypothesis Bayes recovery error.

For the common scoring window, the candidate risk is

$$\mathcal{R}(d, a, b) = \frac{1}{n_{\mathrm{P2}}} \sum_{t \in \mathcal{I}_L^{\mathrm{P2}}} \mathbb{E}^\star[(Y_t - aY_{t-1} - bU_{t-d})^2],$$

with all expectations taken under the true data-generating law.

Local-sequence contract. Statements with $\Delta_{n_{\mathrm{eff}}} \to 0$ and $n_{\mathrm{eff}}\Delta_{n_{\mathrm{eff}}} \to \infty$ describe a sequence of statistical problems indexed by n. Parameters such as α_n or b_n may change with n. For one fixed recoverable stationary system, the per-observation forecast margin need not go to zero as more observations are collected.

Forecast-only selection contract. Training, checkpoint choice, and structural selection do not use the true delay label. Candidate structure is selected by forecast MSE on an independent selection set. The ideal profiled candidate risks are used only after selection to evaluate τ_{total} and S_{total}.

B Proofs

The proof dependency follows the logical hierarchy of the paper:

1. information identities: P1/P2 KL expressions and Bayes recovery quantities;
2. forecast margins: conversion from structural mismatch to population prediction gaps;
3. selection stability: a deterministic perturbation bound and the $2\tau_{\mathrm{total}}$ stability threshold.

The dependency graph is

KL identities $\to$ forecast margin $\to$ near-optimal structure $\to$ structural stability.

Auxiliary statements referenced below are collected in Appendix C; the proof section is kept together so that all derivations appear in one place.

B.1 Proof of Proposition C.1

Both candidates are Gaussian with covariance $\sigma^2 I$ and means

$$\mu_d = \alpha S_d u, \qquad \mu_{d'} = \alpha S_{d'} u.$$

The equal-covariance Gaussian KL formula gives

$$D_{\mathrm{KL}}(P_d^{(u)} \| P_{d'}^{(u)}) = \frac{1}{2}(\mu_d - \mu_{d'})^\top (\sigma^2 I)^{-1}(\mu_d - \mu_{d'}) = \frac{\alpha^2}{2\sigma^2} \|S_d u - S_{d'} u\|_2^2 .$$

B.2 Proof of Corollary C.2

Stationarity gives

$$\mathbb{E}[(U_{t-d} - U_{t-d'})^2] = 2\sigma_U^2[1 - \rho_U(|d - d'|)].$$

Summing over n_{eff} positions and using Theorem C.1 proves the result.

B.3 Proof of Proposition C.3

Let

$$\Delta_M^2 = (\mu_d - \mu_{d'})^\top \Sigma^{-1}(\mu_d - \mu_{d'}).$$

For equal priors and shared covariance, the optimal Gaussian decision rule has error

$$\Phi(-\Delta_M/2).$$

Since

$$D_{\mathrm{KL}}(P_d \| P_{d'}) = \Delta_M^2/2,$$

the result follows.

B.4 Proof of Proposition C.4

The conditional trajectory KL is the sum of one-step KL terms. At time t, the two conditional means are

$$a^\star Y_{t-1} + b^\star u_{t-d}, \qquad a^\star Y_{t-1} + b^\star u_{t-d'}.$$

The shared output term cancels, leaving

$$b^\star(u_{t-d} - u_{t-d'}).$$

Summing the equal-variance Gaussian KL terms gives the result.

B.5 Proof of Proposition 4.1

For the two fixed-mean Gaussian candidates,

$$D_{\mathrm{KL}}(d_0 \| d_1 \mid u) = \frac{\alpha^2}{2\sigma^2} \|S_{d_0} u - S_{d_1} u\|_2^2. \tag{19}$$

The equal-prior Bayes error is

$$e_{\mathrm{rec}}^\star(d_0, d_1; u) = \Phi\left(-\sqrt{D_{\mathrm{KL}}(d_0 \| d_1 \mid u)/2}\right), \tag{20}$$

where Φ is the standard normal cumulative distribution function.

Write $n = n_{\mathrm{eff}}$ and $\Delta_n = \Delta_{n_{\mathrm{eff}}}$ for this proof. By Equations (4) and (19),

$$D_{\mathrm{KL}}(d_0 \| d_1 \mid u) = \frac{n\Delta_n(d_1; d_0, u)}{2\sigma^2}.$$

Substituting into Equation (20) gives the Bayes-recovery-error formula.

Under the true delay,

$$Y = \alpha S_{d_0} u + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n).$$

Therefore

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|Y-\alpha S_{d_1}u\|_2^2 &= \frac{1}{n}\mathbb{E}\|\epsilon+\alpha(S_{d_0}-S_{d_1})u\|_2^2\\ &= \sigma^2+\frac{\alpha^2}{n}\|(S_{d_0}-S_{d_1})u\|_2^2\\ &= \sigma^2+\Delta_n.\end{aligned}$$

If $\Delta_{n_{\text{eff}}} \to 0$ and $n_{\text{eff}}\Delta_{n_{\text{eff}}} \to \infty$, the forecast penalty goes to zero and the Bayes-recovery-error argument goes to $-\infty$.

B.6 Proof of Proposition 4.2

Write $n = n_{\mathrm{P2}}$, $\boldsymbol{x}_{t,d} = (Y_{t-1}, U_{t-d})^\top$, $\boldsymbol{\beta} = (a,b)^\top$, and $m_t^\star = a^\star Y_{t-1} + b^\star U_{t-d^\star}$. All sums below run over $\mathcal{I}_L^{\mathrm{P2}}$, and the boundary/input marginal is shared as specified in Appendix A. For candidate d,

$$Y_t = m_t^\star + \epsilon_t, \qquad \mathbb{E}^\star[\epsilon_t \boldsymbol{x}_{t,d}] = 0.$$

Hence

$$\mathcal{R}_n(d,\boldsymbol{\beta}) = \sigma^2 + \frac{1}{n}\sum_t \mathbb{E}^\star[(m_t^\star - \boldsymbol{x}_{t,d}^\top\boldsymbol{\beta})^2]. \tag{21}$$

Define

$$Q_d := \inf_{\boldsymbol{\beta}} \sum_t \mathbb{E}^\star[(m_t^\star - \boldsymbol{x}_{t,d}^\top\boldsymbol{\beta})^2].$$

Then

$$Q_d = n\Delta_{\mathrm{prof},\mathrm{n_{P2}}}(d).$$

With fixed candidate variance σ^2,

$$D_{\mathrm{KL}}(P^\star \| P_{d,\boldsymbol{\beta},\sigma^2}) = \frac{1}{2\sigma^2}\sum_t \mathbb{E}^\star[(m_t^\star - \boldsymbol{x}_{t,d}^\top\boldsymbol{\beta})^2].$$

Profiling over $\boldsymbol{\beta}$ gives

$$J_{\mathrm{prof}}^{\mathrm{fixed}} = \frac{n\Delta_{\mathrm{prof},\mathrm{n_{P2}}}(d)}{2\sigma^2}.$$

If the candidate uses free shared variance s^2, the KL is

$$\frac{1}{2}\left[n\log\frac{s^2}{\sigma^2} + \frac{n\sigma^2+Q_d}{s^2} - n\right].$$

The minimizer is

$$s_{\mathrm{opt}}^2 = \sigma^2 + \frac{Q_d}{n}.$$

Substitution gives

$$J_{\mathrm{prof}}^{\mathrm{free}} = \frac{n}{2}\log\left(1 + \frac{\Delta_{\mathrm{prof},\mathrm{n_{P2}}}(d)}{\sigma^2}\right).$$

B.7 Derivation of the P2 partial-correlation margin

We work in the centered stationary Gaussian setting with nonsingular predictor covariance. Conditional variances below are therefore constant and coincide with linear projection residual variances.

The conditional-variance form in equation 10 measures the part of the true delayed input that cannot be predicted from the wrong lag and past output. We derive its equivalent partial-correlation representation below.

By definition, after re-fitting the coefficients of the alternative delay d, the remaining prediction error is

$$\Delta_{\text{prof}}(d) = (b^\star)^2 \inf_{a',b'} \mathbb{E}^\star \left[\left(U_{t-d^\star} - a' Y_{t-1} - b' U_{t-d} \right)^2 \right].$$

The linear component explained by Y_{t-1} can be removed from both delayed inputs. Applying the Frisch–Waugh–Lovell theorem, the above optimization is equivalent to first removing the contribution of Y_{t-1} and then projecting the remaining part of $U_{t-d^\star}$ onto the remaining part of U_{t-d}.

Therefore, the residual prediction error can be written directly as

$$\inf_{a',b'} \mathbb{E}^\star \left[\left(U_{t-d^\star} - a' Y_{t-1} - b' U_{t-d} \right)^2 \right] = \operatorname{Var}\left(U_{t-d^\star} \mid Y_{t-1} \right) \left(1 - \rho^2_{U_{t-d^\star}, U_{t-d} \mid Y_{t-1}} \right).$$

Hence,

$$\boxed{\Delta_{\text{prof}}(d) = (b^\star)^2 \operatorname{Var}\left(U_{t-d^\star} \mid Y_{t-1} \right) \left(1 - \rho^2_{U_{t-d^\star}, U_{t-d} \mid Y_{t-1}} \right).}$$

This expression shows why re-fitting reduces the forecast margin. The first factor measures how much of the true delayed input remains unexplained by the past output. The second factor measures how much of this remaining information cannot be replaced by the alternative lag.

B.8 Proof of Proposition C.5

If $\varepsilon < \Delta^{\min}$, every alternative d satisfies

$$r^\star(d) > r^\star(d^\star) + \varepsilon,$$

so the near-optimal set contains only $d^\star$. If $\varepsilon \geq \Delta^{\min}$, an alternative attaining the minimum margin belongs to the near-optimal set.

B.9 Proof of Theorem 4.3

For any alternative d,

$$\begin{aligned} \widehat{r}(d) - \widehat{r}(d^\star) &\geq [r^\star(d) - \tau_{\text{total}}] - [r^\star(d^\star) + \tau_{\text{total}}] \\ &\geq \Delta^{\min} - 2\tau_{\text{total}}. \end{aligned}$$

Thus $\Delta^{\min} > 2\tau_{\text{total}}$ makes every alternative strictly worse.

For sharpness, let d_1 attain the minimum margin. Set

$$\widehat{r}(d^\star) = r^\star(d^\star) + \tau_{\text{total}}, \qquad \widehat{r}(d_1) = r^\star(d_1) - \tau_{\text{total}},$$

and leave the other candidate values unchanged. Then

$$\widehat{r}(d_1) - \widehat{r}(d^\star) = \Delta^{\min} - 2\tau_{\text{total}}.$$

This is negative below the boundary and zero at equality.

C Additional Theoretical Results

Proposition C.1 (Conditional P1 delay information)**.** *For any $d, d' \in \mathcal{D}$,*

$$D_{\text{KL}}(P_d^{(u)} \| P_{d'}^{(u)}) = \frac{\alpha^2}{2\sigma^2} \left\| S_d u - S_{d'} u \right\|_2^2.$$

Corollary C.2 (Average P1 information under stationary input)**.** *If U_t is second-order stationary with variance σ_U^2 and autocorrelation ρ_U, then for $r = |d - d'|$,*

$$\mathbb{E}_U[D_{\text{KL}}(P_d^{(U)} \| P_{d'}^{(U)})] = n_{\text{eff}} \frac{\alpha^2 \sigma_U^2}{\sigma^2} [1 - \rho_U(r)].$$

Proposition C.3 (Exact binary Bayes recovery error)**.** *For two equal-prior Gaussian candidates with the same covariance,*

$$e^{\star}_{\mathrm{rec}}(d, d'; u) = \Phi\left(-\sqrt{D_{\mathrm{KL}}(P^{(u)}_{d} \| P^{(u)}_{d'})/2}\right).$$

Proposition C.4 (Known-parameter P2 delay information)**.** *If the two P2 candidates use the same true coefficients and variance, then*

$$D_{\mathrm{KL}}(P^{(u)}_{d,(a^\star,b^\star)} \| P^{(u)}_{d',(a^\star,b^\star)}) = \frac{(b^\star)^2}{2(\sigma)^2} \sum_{t \in \mathcal{I}_L^{\mathrm{P2}}} (u_{t-d} - u_{t-d'})^2.$$

Proposition C.5 (Near-optimal forecasting and structure)**.** *Let $d^\star$ uniquely minimize the ideal candidate risk and let*

$$\Delta^{\min} = \min_{d \neq d^\star} [r^\star(d) - r^\star(d^\star)].$$

For

$$\mathcal{A}(\varepsilon) = \{d : r^\star(d) \leq r^\star(d^\star) + \varepsilon\},$$

we have

$$\mathcal{A}(\varepsilon) = \{d^\star\} \quad \textit{if} \quad 0 \leq \varepsilon < \Delta^{\min},$$

while $\mathcal{A}(\varepsilon)$ contains an alternative structure when $\varepsilon \geq \Delta^{\min}$.

Proposition C.6 (Partial-correlation form of the profiled margin)**.** *In the centered stationary Gaussian P2 setting with nonsingular predictor covariance, the reduction of the profiled forecast margin can be characterized by*

$$\Delta_{\mathrm{prof}}(d) = (b^\star)^2 \operatorname{Var}(U_{t-d^\star} \mid Y_{t-1}) \left(1 - \rho^2_{U_{t-d^\star}, U_{t-d} | Y_{t-1}}\right).$$

The first factor measures the part of the true delayed input that cannot be explained by the past output. The second factor measures the part of this remaining signal that cannot be replaced by the alternative lag.

This expression explains why re-fitting can reduce the profiled forecast margin. If the past output or another lag already contains predictive information about the true delayed input, the alternative structure can recover part of the missing signal after re-fitting, leaving a smaller forecast difference.

D Bayes Oracle Computation

For a finite set of fully specified candidates with prior masses π_j and conditional densities $p_j^{(u)}$, the Bayes recovery error is

$$e^{\star}_{\mathrm{rec}}(u) = 1 - \int \max_j \{\pi_j p_j^{(u)}(y)\}\, dy.$$

For binary P1, the closed form in Equation (20) is used. In numerical checks, the MAP rule is also evaluated by Monte Carlo using independently generated candidate labels and Gaussian trajectories. The Monte Carlo standard error of an empirical error rate $\widehat{R}$ based on B independent decisions is

$$\sqrt{\widehat{R}(1 - \widehat{R})/B}.$$

E Experimental Protocols

E.1 P1 scale study

The P1 phase study uses

$$\beta \in \{0.5, 1, 1.25\}, \qquad n \in \{128, 256, 512, 1024, 2048, 4096, 8192\},$$

with 200 independent input trajectories per condition and 10,000 MAP decisions per condition. The master seed is fixed. The experiment checks the exact KL identity, analytic/MAP Bayes agreement, and the slopes of Δ_n and $n\Delta_n$.

E.2 P2 POPULATION MARGIN STUDY

The P2 population study uses stationary Gaussian AR(1) input and scans

$$a \in \{0, 0.3, 0.6, 0.85\}, \qquad \rho \in \{0, 0.3, 0.6, 0.85\}, \qquad \sigma \in \{1, 0.5, 0.2\},$$

with $b = 0.5$, true delay 2, and alternative delay 1. Stationary covariances are solved analytically by linear state-space equations. The profiled forecast margin is computed by Schur complement, partial correlation, and direct regression. For state $s_t = (Y_t, U_t, U_{t-1})^\top$, the covariance C solves

$$C = FCF^\top + Q, \qquad F = \begin{pmatrix} a & 0 & b \\ 0 & \rho & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad Q = \mathrm{diag}(\sigma^2, 1 - \rho^2, 0).$$

The prediction feature vector before the transition is $z_t = (Y_{t-1}, U_{t-1}, U_{t-2})^\top$, with stationary covariance C. The true and wrong candidates use coordinates $(1, 3)$ and $(1, 2)$, respectively, in one-based indexing. For $X = (Y_{t-1}, U_{t-1})$, the Schur-complement calculation is

$$\Delta_{\mathrm{prof}} = b^2\{\mathrm{Var}(U_{t-2}) - \mathrm{Cov}(U_{t-2}, X)\,\mathrm{Var}(X)^{-1}\,\mathrm{Cov}(X, U_{t-2})\}.$$

The comparator $\Delta_{\mathrm{lag}} = b^2(1 - \rho^2)$ fixes $a = a^\star$ but optimizes the wrong-delay input coefficient. If both coefficients are instead held at their true values, the adjacent-lag margin is $2b^2(1 - \rho)$, equal to 0.075 in the representative condition of Table 3. These are distinct coefficient constraints.

Numerical check of re-fitting. We use unit-variance Gaussian AR(1) inputs, true delay 2, and alternative delay 1. The lag-only margin $\Delta_{\mathrm{lag}} = b^2(1 - \rho^2)$ fixes $a = a^\star$ and re-fits b, while Δ_{prof} re-fits both coefficients. Table 3 gives their values at $(a, \rho, \sigma, b) = (0.3, 0.85, 0.2, 0.5)$.

Table 3: P2 margins under two coefficient constraints in the same system.

Quantity	Wrong-delay coefficients	Margin
Δ_{lag}	Fix $a = a^\star$, re-fit b	0.069375
Δ_{prof}	Re-fit both a and b	0.046309

The inequality follows from nested optimization; the calculation quantifies the remaining signal in equation 10 and its profiled trajectory KL. Across 48 conditions, Schur-complement, partial-correlation, and direct-regression calculations agree to numerical precision (Appendix E.2).

E.3 CONTROLLED P1 LEARNING STUDY

The controlled P1 learning study keeps accumulated structural information fixed while scanning the per-step margin. A small soft or hard straight-through gate receives only context-based candidate forecast evidence and is trained by forecast MSE. The experiment checks that Bayes/MAP recoverability remains nearly fixed within each information level. Its main result is negative: selection accuracy does not show a stable decrease as the margin becomes smaller.

E.4 FINITE-OBJECTIVE STABILITY STUDY

The finite-objective study uses 90 conditions over forecast margin, objective sample size, and noise level, with $5{,}000$ independent replicates per condition. The ideal candidate risks are known exactly, so τ_{total} is measured directly. A separate deterministic construction checks the sharp boundary at $\Delta = 2\tau$.

The Spearman coefficient in Table 4 summarizes the association between ordered ratio bins and their empirical error rates within this fixed experimental grid. It is a descriptive aggregate result, not the correlation between the ratio and individual selection outcomes, and not a consequence of the sufficient-condition theorem. Its magnitude can depend on the binning and the mixture of experimental conditions. We therefore do not interpret it as a universal difficulty scale or as evidence of predictive superiority on unobserved mechanisms. Appendix G.4 gives a complementary P1 sequence in which recovery improves while the typical ratio decreases.

Table 4: Finite-objective stability study. $S_{\text{total}} > 1$ is the sufficient region from Theorem 4.3.

Quantity	Result
Independent objective samples	450,000
Samples with $S_{\text{total}} > 1$	97,238
Wrong selections with $S_{\text{total}} > 1$	0
Wrong-selection rate with $S_{\text{total}} \leq 1$	19.48%
Descriptive Spearman correlation across fixed S_{total} bins	-0.857

E.5 P1 NEURAL CANDIDATE STUDY

The P1 neural candidate study trains separate candidate-restricted neural forecasters and selects between them using independent forecast MSE. Ideal candidate risks are known analytically. This allows the total objective distortion to be measured and also supports a finer decomposition into trained-model and finite-selection contributions.

E.6 P2 LEARNED-CANDIDATE PROTOCOL

The 12-mechanism grid uses $a \in \{0, 0.6\}$, $\rho \in \{0, 0.6, 0.9\}$, $\sigma \in \{0.5, 2\}$, and $b = 0.5$, with training sizes $64, 256, 1024$ and selection sizes $32, 128, 512$. It contains 288 fitted candidate pairs per learner and 864 selections. The sufficient region contains 319 OLS and 246 MLP selections. The true delay is 2 and the alternative delay is 1, with

$$U_t = \rho U_{t-1} + \sqrt{1-\rho^2}\,\eta_t, \qquad Y_t = aY_{t-1} + bU_{t-2} + \sigma\epsilon_t,$$

where η_t, ϵ_t are independent standard Gaussian innovations. Each of the eight independent blocks per mechanism generates one training trajectory of length 1024 and an independent selection trajectory of length 512, initialized from the stationary covariance in Appendix E.2. Smaller sizes use their common prefixes. Initial states and subsequent innovations use independent random streams; population audits are independent of both trajectories. Thus the training and selection observations are continuous ARX transitions, rather than independent draws from a stationary one-step marginal.

OLS includes an intercept. The MLP has widths 2–16–16–1 with GELU activations and is fitted for 400 AdamW updates (learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, gradient clipping at 1). Each candidate uses input and target standardization computed only from its own training prefix, with population-form standard deviations (divisor n). The final checkpoint is used without selection-based early stopping. All selection and population risks are reported in the original target-squared units. Exact empirical ties use the fixed candidate priority $[2, 1]$; the selector does not read the true delay or the population audit.

Population risk of frozen predictors. Let $R_d = \mathbb{E}^\star[(Y - \widehat{f}_d(X_d))^2]$, conditional on the fitted function. For OLS, this is evaluated analytically from C. For an intercept c_d and a three-coordinate coefficient vector θ_d with a zero at the excluded input lag,

$$R_d = \sigma^2 + c_d^2 + (\theta^\star - \theta_d)^\top C(\theta^\star - \theta_d), \qquad \theta^\star = (a, 0, b)^\top.$$

MLP population ordering is audited using $65{,}536$ independent draws of stationary z. Writing $\mu(z) = az_1 + bz_3$, we average the paired squared-error difference

$$T(z) = (\mu(z) - \widehat{f}_1(z_1, z_2))^2 - (\mu(z) - \widehat{f}_2(z_1, z_3))^2.$$

Independent marginal draws are appropriate for this population integral, although they are not used as training or selection trajectories. A per-pair 99% Monte Carlo interval is the sample mean plus or minus 2.575829 sample standard errors. An interval crossing zero yields an unresolved population ordering. These are integration-error intervals for fixed predictors, not simultaneous recovery guarantees.

Risk levels. For true and wrong candidates t, w, define

$$g_{\text{train}} = (R_w - r_w^\star) - (R_t - r_t^\star),$$

$$g_{\mathrm{pop}} = R_w - R_t = \Delta + g_{\mathrm{train}},$$
$$g_{\mathrm{eval}} = (\widehat{r}_w - R_w) - (\widehat{r}_t - R_t),$$
$$\widehat{g} = \widehat{r}_w - \widehat{r}_t = \Delta + g_{\mathrm{train}} + g_{\mathrm{eval}}.$$

These identities distinguish the ideal margin, the ordering of frozen predictors, and the empirical selection outcome. A population inversion is not attributed solely to optimization error: finite training, the predictor class, and the fitting procedure all enter $R_d - r_d^\star$.

E.7 Why forecast margin alone is insufficient

We use a controlled P1 learner in which accumulated recovery information is kept nearly fixed while the per-observation forecast margin varies. A small gate receives candidate forecast evidence and is trained only through forecast MSE. The initial gradient decreases approximately proportionally to Δ (log–log slopes 0.987 for the soft gate and 1.01 for the hard straight-through gate), but structural accuracy does not exhibit a corresponding monotone deterioration. This is a statement about the observed gradient scale, not a general theorem that smaller margins make optimization harder. Together with Proposition 4.1, it shows why the per-observation margin alone does not determine structural recovery or selection failure.

E.8 Data-size effects on empirical selection

We compare training sizes $64 \to 1024$ at fixed selection size and selection sizes $32 \to 512$ at fixed training size. Table 5 reports the difference in selection error rate (larger minus smaller size), averaged equally across the fixed 12 mechanisms. We obtain 95% percentile intervals from $2{,}000$ bootstrap draws, resampling the eight independent blocks within each mechanism and preserving the candidate, learner, and size pairing within each block. Mechanisms are resampled independently, with analysis seed 20260908. Individual transitions and the three selection prefixes are not treated as independent replicates.

Table 5: Paired endpoint differences in structural-selection error rate. Negative values indicate improvement. The fixed size refers to the axis that is not varied.

Learner	Varied axis	Fixed size	Difference	95% interval
OLS	Training	32	−0.0417	[−0.1042, 0.0208]
OLS	Training	128	−0.0208	[−0.0625, 0.0208]
OLS	Training	512	−0.0729	[−0.1250, −0.0313]
OLS	Selection	64	−0.0833	[−0.1563, −0.0104]
OLS	Selection	256	−0.1042	[−0.1667, −0.0417]
OLS	Selection	1024	−0.1146	[−0.1771, −0.0625]
MLP	Training	32	−0.0729	[−0.1667, 0.0313]
MLP	Training	128	−0.0208	[−0.1042, 0.0625]
MLP	Training	512	−0.1354	[−0.1979, −0.0729]
MLP	Selection	64	−0.0729	[−0.1563, 0.0000]
MLP	Selection	256	−0.0417	[−0.1146, 0.0313]
MLP	Selection	1024	−0.1354	[−0.1979, −0.0833]

Increasing selection size for frozen predictors changes the empirical difference but leaves their population ordering unchanged. Increasing training size changes the fitted predictors and their population ordering, while using the same selection trajectory within a block. The MLP training contrast is small and uncertain at selection size 128, but larger at selection size 512. Nine individual mechanism/axis/fixed-size contrasts have point estimates opposite to the corresponding pooled direction; their intervals all contain zero. Thus neither a pooled improvement nor these individual reversals establishes a universal monotone data-size effect.

The conclusions are limited to these mechanisms, eight blocks, and the fixed 400-update MLP budget. Resampling-based analyses involving audited population quantities condition on the fixed Monte Carlo audit rather than propagating its integration uncertainty. With eight blocks, zero observed errors or a degenerate bootstrap interval cannot establish zero population risk or an exactly absent effect.

F SEMI-SYNTHETIC INDUSTRIAL DYNAMICS WITH CONTROLLED TEMPORAL STRUCTURE

This appendix provides a semi-synthetic illustration using realistic industrial process dynamics with a controlled temporal structure. It is not a recovery benchmark for physical process delays; the structural label is injected by construction, while the temporal background is inherited from the Tennessee Eastman Process (TEP) simulator.

We use normal-operation TEP trajectories as background dynamics and construct semi-synthetic targets by injecting a known delay,

$$Y_t^{\text{new}} = aY_{t-1}^{\text{new}} + bX_{t-d^\star} + \sigma\epsilon_t.$$

The injected delay $d^\star$ is used only after forecast-only selection for evaluation. For each task, candidate delays $d \in \{0, \ldots, 20\}$ are evaluated by the same forecast-only protocol as in the main experiments: a separate predictor is trained for each candidate delay, and the selected structure is $\hat{d} = \arg\min_d \widehat{r}(d)$ on independent forecast MSE.

The study varies injected delays, noise levels, training size, and selection size. Figure 5 shows a representative candidate-risk profile, recovery–margin comparison, and stability illustration. The results are intended only as an applied sanity check: smaller per-observation forecast differences can coexist with similar or improved selection accuracy, while finite-data errors increase when the forecast margin becomes harder to distinguish.

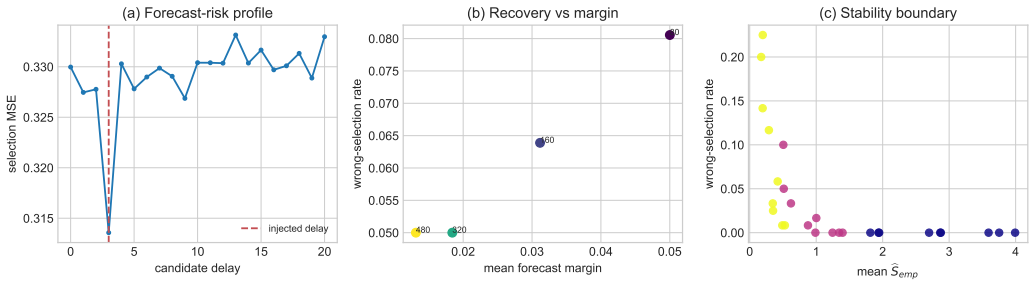


Figure 5: Semi-synthetic industrial dynamics experiment. The experiment uses Tennessee Eastman Process trajectories as realistic temporal backgrounds while injecting controlled temporal structures. The panels show (a) the forecast risk profile over candidate delays, (b) representative structural selection behavior under repeated evaluation, and (c) the relationship between forecast margin and selection consistency.

G ADDITIONAL EMPIRICAL RESULTS

G.1 P1 NEURAL ERROR SOURCES

In the P1 neural candidate study, 86.02% of selections are fully correct. Model-level candidate ordering is reversed in 2.82% of cases that remain wrong after finite selection, while 9.95% are finite-selection flips from a correct trained-model ordering. Another 1.20% are accidental corrections of a model-level inversion. The stability guarantee holds in every case with $S_{\text{total}} > 1$.

G.2 P2 POPULATION ORDERING AND EMPIRICAL SELECTION

Population ordering is counted once per fitted candidate pair; empirical outcomes are counted at each selection size. Of 288 OLS pairs, 276 favor the true delay and 12 favor the wrong delay. For MLP, these counts are 271 and 16, with one unresolved pair. The unresolved pair has $a = 0.6$, $\rho = 0.6$, $\sigma = 2$, training size 64, and block 5; its estimated wrong-minus-true population gap is 0.0054401, with 99% Monte Carlo interval $[-0.0065748, 0.017455]$.

The joint outcomes are summarized in Table 2 in the main text. The three unresolved MLP selections share one fitted pair.

A correct empirical choice from a population-inverted pair is a finite-sample outcome; it does not imply that less selection data is preferable. These counts also distinguish population-ordering uncertainty in one MLP pair from errors caused by empirical selection.

G.3 Auxiliary correlation results under independent marginal sampling

As an auxiliary check, we also ran a learned-candidate experiment with independent Gaussian input–output transitions drawn from the stationary P2 one-step marginal. This is a separate sampling protocol from the continuous-trajectory experiments and is not used to establish calibration or transfer. The grid uses $a \in \{0, 0.6, 0.85\}$, $\rho \in \{0, 0.6, 0.85, 0.95\}$, $\sigma \in \{1, 2, 4\}$, $b \in \{0.5, 0.25\}$, training sizes $16, 64, 256$, selection sizes $8, 32, 128$, and eight training seeds, giving $1{,}728$ fitted candidate pairs and $5{,}184$ selection outcomes. Within this pooled protocol, grouped error rate has Spearman correlation -0.905 with S_{total}; the individual-outcome point-biserial correlations are -0.0817 for $\log \Delta_{\mathrm{prof}}$ and -0.1975 for $\log S_{\mathrm{total}}$. These are descriptive auxiliary observations only, and pooled associations can depend on the grouping and mixture of mechanisms.

G.4 Scope of descriptive ratio trends

Adding the same constant c to every empirical objective leaves the selected structure unchanged, while it can change τ_{total} and hence S_{total}. Thus a descriptive ratio trend within a specified experiment is not a universal difficulty law; this conservativeness is compatible with the sufficient theorem.

The negative binned correlation in Table 4 is compatible with opposite trends along other problem sequences. In an additional fixed-input P1 sequence with $\beta = 0.75$, $\sigma = 1$, and $\Delta_n = n^{-\beta}$, the exact recovery parameter is $Q = \sqrt{n\Delta_n}/(2\sigma)$. Each sample size has $20{,}000$ independent Gaussian-noise replicates. The sequence gives the following results:

Table 6: A P1 local sequence illustrating the scope of the ratio. Recovery error decreases while the median oracle S_{total} also decreases; the sufficient theorem is unaffected.

n	Empirical recovery error	Exact recovery error	Median S_{total}
256	0.15845	0.15866	0.11835
1024	0.11905	0.11718	0.087417
4096	0.07860	0.07865	0.064037
16384	0.04585	0.046305	0.045355
65536	0.02235	0.02275	0.032723
262144	0.00990	0.0086936	0.022668

For two fixed candidate means separated by $\|\mu_t - \mu_w\|^2 = n\Delta_n$, the experiment samples the exact joint law of the empirical losses through independent $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_{n-1}$:

$$A = \frac{\sigma^2}{n}(Z^2 + W - n), \quad B = 2\sigma\sqrt{\Delta_n/n}\, Z, \quad \widehat{r}_t = \sigma^2 + A, \quad \widehat{r}_w = \sigma^2 + \Delta_n + A + B.$$

Ideal risks are σ^2 and $\sigma^2 + \Delta_n$, and $\tau_{\mathrm{total}} = \max\{|A|, |A+B|\}$. This exact loss simulation preserves the dependence between the two candidate losses. Shared fluctuations in the objective can affect τ_{total} while cancelling in the candidate difference. The theorem bounds every candidate's absolute objective distortion, so it can be conservative for the actual comparison. These results motivate reporting a within-experiment correlation as an auxiliary observation rather than assigning S_{total} a universal monotone interpretation.

G.5 Predictive substitution and selection

Predictive substitution determines the profiled population margin through equation 10. The sufficient selection condition compares that margin with distortion in the empirical objective. A change in input dependence or output memory may affect both quantities, so neither the population identity nor the sufficient theorem alone predicts a monotone change in empirical selection error.

H COMPUTABLE STRUCTURAL CERTIFICATION: CONSTRUCTION AND GUARANTEE

This section spells out the conditional split/profile likelihood application in Section 4.4. The underlying method is universal inference (Wasserman et al., 2020); it does not estimate the oracle stability ratio and does not supply a new universal law of selection difficulty.

H.1 CONDITIONAL TRAJECTORY CONSTRUCTION

Let A, B, C be independent training, selection, and certification trajectories. The candidate family and the certification length are fixed before evaluating outcomes in C. Training uses A, and the forecast-only proposal $\widehat{d}$ minimizes selection MSE on B. Before certification, freeze the proposal, its fitted mean function $\widehat{f}_{\widehat{d}}$, and a strictly positive variance estimate $\widehat{v}$ obtained without C. Write n_C for the number of scored certification transitions. Conditional on its initial history and observed exogenous input trajectory, the frozen predictor defines

$$Q(C) = \prod_{t=1}^{n_C} \varphi_{\widehat{v}}\Big(Y_t - \widehat{f}_{\widehat{d}}(Y_{t-1}, U_{t-\widehat{d}})\Big), \tag{22}$$

where φ_v is the centered Gaussian density of variance v. A normalized sequential conditional density remains valid with random output histories: the product in equation 22 is a density in the entire output trajectory. The argument does *not* condition on the full design matrix containing future output histories as if it were fixed independent data.

For candidate d, the structural family has conditional likelihood

$$p_{d,\theta}(C) = \prod_{t=1}^{n_C} \varphi_{s^2}(Y_t - c - aY_{t-1} - bU_{t-d}), \qquad \theta = (c, a, b, s^2), \quad s^2 > 0. \tag{23}$$

The intercept enlarges the centered P2 family without excluding its true distribution. All likelihoods condition on the same exogenous inputs and boundary history; no candidate-specific stationary initial density is multiplied into them. Stationarity is used to initialize simulations, not to prove the conditional likelihood guarantee.

The certificate uses E_d and $\mathcal{C}_\alpha$ from equation 16. A retained candidate means that its structural family has not been rejected by this test; it is not a posterior probability or a claim that the candidate is true. An empty set, multiple retained candidates, or a singleton different from $\widehat{d}$ leads to abstention. Certification does not change the original forecast-only answer.

H.2 PROOF OF PROPOSITION 4.4

Let $\mathcal{F}$ contain all pre-certification information together with the certification trajectory's exogenous input and initial history. Under the stated exogeneity and splitting conditions, the true conditional density is some $p_{d^\star,\theta^\star}$. Pointwise,

$$E_{d^\star} \le \frac{Q(C)}{p_{d^\star,\theta^\star}(C)}.$$

Since Q is normalized and fixed given $\mathcal{F}$,

$$\mathbb{E}^\star[E_{d^\star} \mid \mathcal{F}] \le \int Q(y_C \mid \mathcal{F})\, \mathrm{d}y_C = 1.$$

Markov's inequality therefore gives

$$\Pr(d^\star \notin \mathcal{C}_\alpha \mid \mathcal{F}) = \Pr(E_{d^\star} \ge 1/\alpha \mid \mathcal{F}) \le \alpha.$$

Integrating over $\mathcal{F}$ proves the confidence-set statement. A wrong certified singleton must exclude $d^\star$, so its probability is also at most α. No candidate-count correction is required for coverage of one fixed true label; this is not simultaneous control of exclusions of every observationally equivalent label.

Joint versus conditional risk. The guarantee controls $\Pr(\widehat{d} \neq d^\star, \text{certified})$. It does not establish $\Pr(\widehat{d} \neq d^\star \mid \text{certified}) \leq \alpha$. If the certification probability is $c_0 > 0$, the direct implication is only an upper bound $\min\{1, \alpha/c_0\}$ on that conditional risk. Certification rate, confidence-set coverage, and error among certified choices are different quantities and are reported separately.

H.3 Closed-form calculation and numerical safeguards

Define the frozen-predictor residual sum of squares and the optimized candidate residual sum of squares by

$$\mathrm{RSS}_Q = \sum_{t=1}^{n_C} \big(Y_t - \widehat{f}_{\widehat{d}}(Y_{t-1}, U_{t-\widehat{d}})\big)^2,$$
$$\mathrm{RSS}_d = \inf_{c,a,b} \sum_{t=1}^{n_C} (Y_t - c - aY_{t-1} - bU_{t-d})^2.$$

For $\mathrm{RSS}_d > 0$, variance profiling gives $\widehat{s}_d^2 = \mathrm{RSS}_d / n_C$ and

$$\log E_d = \frac{n_C}{2}\left[\log \frac{\mathrm{RSS}_d / n_C}{\widehat{v}} + 1\right] - \frac{\mathrm{RSS}_Q}{2\widehat{v}}. \tag{24}$$

The implementation computes candidate least-squares fits by SVD and compares log ratios with $\log(1/\alpha)$. Fitting the denominator on C is part of the prespecified statistic, not training-data leakage. Suspected denominator degeneracy is handled conservatively by retaining the candidate; nonfinite predictor outputs raise an error rather than silently deleting a candidate. Exact zero residual permits an unbounded Gaussian null likelihood and is not evidence for rejecting that null.

The denominator must be the actual supremum or a valid upper bound. A lower value found by local neural optimization can inflate E_d and invalidate the guarantee. Neural predictors are used only in the numerator in the present experiments. If an OLS numerator belongs to its selected null family, its selected-candidate ratio cannot exceed one; this containment argument must not be assumed for an MLP numerator.

H.4 Limits of interpretation

Validity requires that the true conditional law belong to the specified candidate family, a normalized pre-certification numerator, correct boundary conditioning, and the declared exogeneity and data-isolation assumptions. Missing true candidates, misspecified transitions, or uncontrolled reuse of certification outcomes are not covered. The reported experiments do not certify arbitrary nonlinear structures or a published temporal-mask explainer. A fixed-length guarantee is not permission to inspect successive prefixes and stop at the first passing certificate.

When $b = 0$, all delay labels can represent the same mechanism. More generally, nested support families may contain the same conditional law. Unique physical or minimal support recovery then requires an explicitly identifiable target; a singleton cannot simply be equated with recovery of an arbitrarily assigned label. These limitations are separate from the oracle condition's conservativeness.

I Computable Certification: Experimental Protocol and Full Results

I.1 Generation, fitting, and data isolation

The formal run, identified in the released report as `certificate_61ded05bf7a1083c`, uses unit-variance AR(1) inputs and continuous P2 trajectories,

$$U_t = \rho U_{t-1} + \sqrt{1-\rho^2}\,\eta_t, \qquad Y_t = aY_{t-1} + bU_{t-2} + \sigma\epsilon_t,$$

with independent standard Gaussian innovations. The five candidates are $\{1, 2, 3, 4, 5\}$. Training, selection, and certification trajectories have separate random streams, each is initialized from the stationary law, and transitions are then generated continuously. All candidates share their scoring

window and use only Y_{t-1} and the permitted lagged input. Structure labels and population quantities are restricted to scoring metadata and are not inputs to the certification decision.

The OLS grid has $a \in \{0, 0.6\}$, $\rho \in \{0, 0.9\}$, $\sigma \in \{0.5, 2\}$, and $b = 0.5$: eight mechanisms with 300 independent blocks each. Training lengths are $64, 256, 1024$, selection length is 128, and certification lengths are $64, 256, 1024$. The 21,600 outcome rows share blocks and prefixes and are not 21,600 independent replications.

The MLP grid fixes $a = 0.6$, with the same two correlations and noise levels and $b = 0.5$: four mechanisms with 40 independent blocks each. Five separate 2–16–16–1 networks are trained per block, giving 800 networks, each for the complete 400 AdamW updates. Training length is 256, selection length is 128, and certification lengths are $128, 512, 2048$. The variance estimate and predictor are frozen before certification. MLPs provide only the numerator; the null families remain the Gaussian linear families in equation 23.

The main threshold is $\alpha = 0.05$, fixed before the run. Prespecified sensitivity thresholds reuse the saved log ratios and do not create additional independent samples. The following tables report the formal run rather than the earlier feasibility pilot; all original candidate predictions and fitted-model hashes are retained by the experiment implementation.

I.2 Certification rates and uncertainty

Certification means that the retained set is exactly the singleton containing the original selection. Table 7 summarizes endpoints and difficult conditions; Table 8 reports every main-threshold training/certification combination. N is the number of independent pipeline blocks, W_0 the original wrong choices, N_{cert} the certified choices, and W_{cert} the erroneous certifications. The denominator of erroneous-certification probability is N, whereas error among certified choices uses N_{cert} and is undefined when $N_{\text{cert}} = 0$.

Table 7: **Certification at the longest prespecified length.** Training length $n_A = 256$, selection length $n_B = 128$, and $\alpha = 0.05$. Entries are counts over N independent blocks. The difficult-condition rows are subsets of the pooled rows, not additional replications.

Setting	Learner	n_C	N	Original errors	Certified	Erroneous certifications
All 8 mechanisms	OLS	1024	2400	418	1686	0
All 4 mechanisms	MLP	2048	160	36	100	0
High correlation and noise: $\rho = 0.9$, $\sigma = 2$						
$a = 0$	OLS	1024	300	151	11	0
$a = 0.6$	OLS	1024	300	169	8	0
$a = 0.6$	MLP	2048	40	24	0	0

Table 8: Formal computable-certification results at $\alpha = 0.05$. Rows reuse training and certification prefixes within blocks. No main-threshold row has an observed erroneous certification.

Learner	n_A	n_C	N	W_0	N_{cert}	Rate (%)	W_{cert}
OLS	64	64	2400	540	686	28.58	0
OLS	64	256	2400	540	1097	45.71	0
OLS	64	1024	2400	540	1313	54.71	0
OLS	256	64	2400	418	784	32.67	0
OLS	256	256	2400	418	1420	59.17	0
OLS	256	1024	2400	418	1686	70.25	0
OLS	1024	64	2400	393	824	34.33	0
OLS	1024	256	2400	393	1514	63.08	0
OLS	1024	1024	2400	393	1773	73.88	0
MLP	256	128	160	36	63	39.38	0
MLP	256	512	160	36	89	55.63	0
MLP	256	2048	160	36	100	62.50	0

Figure 4(a) and Table 9 use 2,000 bootstrap resamples of complete blocks within each mechanism, preserving all paired training/certification prefixes and threshold reuse. Mechanisms are weighted equally. They do not treat individual transitions or repeated prefixes as independent observations.

Table 9: Certification rates at $n_A = 256$. Intervals are mechanism-stratified block-bootstrap 95% intervals; the OLS and MLP mechanism grids differ.

Learner	n_C	Rate (%)	95% interval (%)
OLS	64	32.67	$[31.67, 33.67]$
OLS	256	59.17	$[58.00, 60.29]$
OLS	1024	70.25	$[69.33, 71.17]$
MLP	128	39.38	$[35.00, 43.75]$
MLP	512	55.63	$[51.25, 60.00]$
MLP	2048	62.50	$[58.13, 66.89]$

Zero observed errors do not establish zero population risk. Within one fixed mechanism/configuration, a one-sided 95% binomial upper bound after zero errors in N blocks is $1 - 0.05^{1/N}$: about 0.994% for $N = 300$, 7.216% for $N = 40$, and 0.299% for $N = 1000$. For the equally weighted aggregate over M mechanisms, the report allocates confidence failure $0.05/M$ to each mechanism and averages the resulting upper bounds. At each fixed $n_A = 256$ endpoint these conservative aggregate erroneous-certification upper bounds are 1.6775% for OLS and 10.3763% for MLP. They are not simultaneous guarantees over the whole grid or bounds on error conditional on certification.

I.3 Mechanism-level behavior

Table 10 retains every mechanism at $n_A = 256$ and the longest prespecified certification length (1024 for OLS, 2048 for MLP). The near-zero rates in the high-correlation, high-noise conditions are not excluded from the aggregate. They show conservativeness and limited power of this procedure in those settings, not a proof that no method can recover the structure.

Table 10: Mechanism-level certification at $n_A = 256$ and each learner's longest certification length.

Learner	a	ρ	σ	N	W_0	N_{cert}	Rate (%)
OLS	0	0	0.5	300	0	300	100.00
OLS	0	0	2	300	41	240	80.00
OLS	0	0.9	0.5	300	6	294	98.00
OLS	0	0.9	2	300	151	11	3.67
OLS	0.6	0	0.5	300	0	300	100.00
OLS	0.6	0	2	300	44	240	80.00
OLS	0.6	0.9	0.5	300	7	293	97.67
OLS	0.6	0.9	2	300	169	8	2.67
MLP	0.6	0	0.5	40	0	40	100.00
MLP	0.6	0	2	40	10	26	65.00
MLP	0.6	0.9	0.5	40	2	34	85.00
MLP	0.6	0.9	2	40	24	0	0.00

I.4 Prediction comparisons and equal evaluation budgets

We compare the structural certificate with two predictive baselines. The predictive comparator keeps the B-selected candidate fixed and compares paired losses on independent C using a one-sided normal approximation with Bartlett/Newey–West HAC covariance, lag $\lfloor n_C^{1/3} \rfloor$, and a Bonferroni adjustment over the $k - 1$ rivals. This is an approximate frozen-predictor risk comparison, not a complete Model Confidence Set procedure and not a finite-sample structural certificate. The same-evaluation-budget baseline freezes the A-trained models and reselects using the sample-count-weighted MSE over B and C; it uses the same total evaluation observations but does not abstain or supply a certification guarantee.

These comparisons are reported only to clarify the role of the structural certificate. At $n_A = 64, n_C = 1024$, the normally fitted OLS predictive comparator accepts $1{,}626$ answers, 63 of which have the wrong structure, while structural certification accepts $1{,}313$ and has no observed erroneous certifications. At the longest MLP length, HAC accepts 107 and structural certification accepts 100, both without observed structural errors. Because the acceptance rates and guarantees differ, these counts do not establish empirical dominance of the structural procedure or imply that all benefit comes from certification rather than additional data.

I.5 Deliberate undertraining counterexample

The counterexample intentionally separates predictor quality from structural truth. Let $Y_t = U_{t-2} + \epsilon_t$, with unit input variance, adjacent-input correlation 0.9, and noise variance one. Freeze the true-delay predictor at $\widehat{f}_2 = 0.2U_{t-2}$ and the wrong-delay predictor at $\widehat{f}_1 = 0.9U_{t-1}$. Their population risks are

$$R(\widehat{f}_2) = 1 + (1 - 0.2)^2 = 1.64, \qquad R(\widehat{f}_1) = 1 + (1 - 0.9^2) = 1.19.$$

The wrong frozen predictor genuinely has lower risk, although the ideal true-delay family has risk one. Across $1{,}000$ independent blocks, the original selection chooses the wrong delay in all blocks, and independent HAC comparison supports that predictor in all blocks. The structural confidence set retains both candidates in every block, producing zero certifications, not corrected point estimates (Figure 4(b)). This is a designed counterexample, not an estimate of normal-training failure frequency.

I.6 Nonidentifiability and implementation checks

In $1{,}000$ additional independent three-trajectory blocks with $b = 0$, all five labels can represent the same mechanism. Every confidence set retains all five candidates; there are no singleton certifications. No unique true label is assigned, so raw structural error and true-label exclusion are reported as undefined rather than scored against an arbitrary label.

The formal report records 27 targeted checks, including explicit Gaussian-density agreement, the OLS numerator-containment property, threshold boundaries, degeneracy, separation of scoring metadata, continuous indexing, covariance, prefix handling, and the predictive comparator. Saved observations, frozen predictions, variances, log ratios, retained sets, and comparator decisions were recomputed in the experiment audit. No main-threshold row triggered a numerical anomaly, exact selection tie, or variance-protection event; these are row-level diagnostics, not additional independent trials. Full code, checkpoints, and audit manifests belong to the experiment release; this manuscript integration does not constitute a new experimental run.