AUDITING NODE-IMPORTANCE MEASURES ON LEARNED DYNAMICAL MODELS WITH EMBEDDED GROUND-TRUTH PROBES

Anonymous authors
Paper under double-blind review

ABSTRACT

Node-importance measures derived from learned dynamical models, including controllability energies, trajectory-level leverage, and attribution through time, are increasingly reported as substantive findings, even though the true influence of any node is unknown in real data. We introduce an audit for such measures based on an embedded ground-truth probe: a small synthetic subsystem added to the analyst's panel and estimated jointly with it using the same model. Because the probe's true multi-step importance is known exactly, the fitted measure can be tested against ground truth under the same estimation conditions and computational pipeline as the real data. A simple diagnostic, the coupling-to-persistence ratio, indicates when the main failure mode is likely to arise, and a theoretical decomposition explains its source. We apply the audit to an existing research pipeline across five empirical panels. The results reveal both a successful validation and three distinct failure modes: the measure succeeds when the estimator can represent the relevant dynamics, inherits the estimator's structural blind spots, becomes dominated by node persistence when coupling is weak, and can be distorted by pooled linearizations evaluated away from the observed trajectory. The audit requires only a few additional channels and no specialized tooling.

1 INTRODUCTION

Learned dynamical models are increasingly used not only to forecast but also to explain: which components drive a system, which are peripheral, and where an intervention would propagate. Such claims often rely on node-importance measures computed from fitted models, including average controllability from Gramians, modal controllability, Granger-style out-strengths, integrated gradients through rollouts, and trajectory-level leverage measures that propagate perturbations through sequences of fitted Jacobians. Across applications from macroeconomics to political development, these measures are interpreted as properties of the systems being studied.

Their central validation problem is simple: in real data, the true multi-step influence of any node is unknown. Researchers therefore rely on indirect checks such as face-valid rankings, stability across random seeds, and agreement with other fitted statistics. These checks can establish reproducibility or plausibility, but they cannot establish that a ranking measures influence. As we show, all three can hold even when it does not. The problem is not merely hypothetical. We audit an existing research pipeline whose trajectory-leverage rankings reproduce nearly exactly across seeds and sessions. Yet on its primary panel, the ranking largely follows own-persistence. An entire class of interactions is also invisible because the estimator cannot represent it, while a previously used alignment statistic reproduces the persistence relationship through a diagonal-inclusive definition.

We make this validation problem empirically testable without replacing the analyst's data, estimator, or downstream importance measure. Our first contribution is an **embedded ground-truth probe**: a six-channel synthetic subsystem appended to the empirical panel and estimated jointly with it. The probe's true trajectory-level importance is exactly computable using the same standardized gauge, companion convention, and window code path as the fitted measure. Equal own-lags remove persistence as a confound within the probe. An exactly tied triple provides a per-fit noise floor, while

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.

paired experimental arms distinguish dynamics that the estimator cannot represent from dynamics that it can represent but fails to recover.

Our second contribution is a **fitted-model diagnostic** that requires no known ground truth. The coupling-to-persistence ratio measures the strength of estimated cross-node coupling relative to own-persistence. Proposition 1 provides a decomposition showing why this distinction matters: persistence contributes directly to horizon-propagated energy, while the contribution of cross-node coupling depends on its magnitude and propagation through the fitted dynamics. As coupling weakens relative to persistence, node rankings can therefore become increasingly dominated by persistence rather than cross-node influence.

Third, we conduct an **audited case study** spanning five empirical panels, two estimator classes, and approximately 800 fits, with bit-exact reproduction of the underlying pipeline. The audit identifies one successful validation and three distinct failure mechanisms. Representational closure is inherited from the estimator, which we demonstrate through an estimator swap. Persistence weighting is isolated by changing a driver's persistence while holding its causal effect fixed. Finally, pooled linearization can evaluate dynamics away from the observed trajectory, producing errors in opposite directions under the two estimator classes. Together, these experiments separate failures of the importance measure from failures inherited from the model on which it is computed.

Fourth, we translate these results into **practice**. We propose a simple audit protocol in which fitted importance rankings are interpreted only after examining the coupling-to-persistence diagnostic and running the embedded probe under the analyst's own data composition and estimator. The probe requires only a few additional channels and uses the existing estimation and importance-measure pipeline.

All results in this paper are reproducible from the released artifacts. Every reported number was recomputed from those artifacts by an automated verifier before submission.

2 Setup and Measures

We first define the fitted dynamical systems and the importance measures that the audit will evaluate. We observe a panel $\{x_t^{(u)} \in \mathbb{R}^N\}$ over units u and times t. Each channel is standardized on the fit window, and all quantities below are defined in this standardized gauge. An estimator fits a one-step map $\hat{x}_{t+1} = f(x_t, \ldots, x_{t-K+1})$. We study two estimator classes used in practice. The first is an *additive* neural Granger VAR, $f_i = \sum_j g_{ij}(x_{j,t}, \ldots, x_{j,t-K+1})$ with per-input subnetworks and ℓ_1 regularization. The second is a *joint* per-target MLP over the full lag window. Both yield Jacobians along the observed trajectory, with $J_t = \partial \hat{x}_{t+1}/\partial x_t$, which we assemble in companion form over the K lags.

The measures we audit quantify how a perturbation to one node propagates through the fitted dynamics. Let $\Phi(t, \tau)$ denote the ordered product of companion Jacobians that propagates a perturbation introduced at time t forward by τ steps. For node j, the resulting window energy is

$$E_j(t; T) = \sum_{\tau=0}^{T-1} \|\Phi(t, \tau) e_j\|^2$$

We normalize these energies to shares within each window and average the shares across windows to obtain trajectory-level scores.

For comparison, static reductions replace the sequence of Jacobians with powers of a single matrix. The *realized* reduction uses the time-averaged Jacobian, while the *pooled* reduction uses a single linearization at the mean state. We also compare the trajectory-level measure with one-step out-strength from the fitted Jacobians ("granger-from-fit"), modal controllability of the averaged system, Jacobian-graph out-degree, integrated gradients through the rollout, and within-unit variance. Precise conventions, including window starts, retention of $\tau = 0$, and tie-breaking, follow the audited pipeline and are fixed in App. A. Ground-truth and fitted importance are computed under exactly the same conventions, so discrepancies cannot be attributed to differences in implementation or normalization.

2.1 DECOMPOSING TRAJECTORY ENERGY INTO PERSISTENCE AND COUPLING

Proposition 1 below isolates a persistence-only contribution to trajectory energy and bounds everything introduced by cross-node coupling. When coupling is small relative to the separation between nodes' persistence terms, the trajectory-energy ranking is therefore determined by persistence.

Write the one-step Jacobian along the observed trajectory as $J_t = D_t + C_t$, where D_t contains the diagonal own-lag terms and C_t the off-diagonal coupling terms, and set $\delta = \max_t \max_i |(D_t)_{ii}|$ and $\gamma = \max_t \|C_t\|_2$. For horizon T, the raw trajectory energy of node j is $E_j(t;T) = \sum_{\tau=0}^{T-1} \|\Phi(t,\tau)\, e_j\|^2$, where $\Phi(t,\tau) = J_{t+\tau-1} \cdots J_t$ and $\Phi(t,0) = I$.

Proposition 1 (Persistence term and coupling remainder)**.** *Let $p_j(t,\tau) = \prod_{s=0}^{\tau-1} (D_{t+s})_{jj}$, with $p_j(t,0) = 1$, and define the own-persistence contribution $P_j(t;T) = \sum_{\tau=0}^{T-1} p_j(t,\tau)^2$. Then*

$$E_j(t;T) \;=\; P_j(t;T) \;+\; R_j(t;T),$$

$$|R_j(t;T)| \;\le\; \sum_{\tau=0}^{T-1} \Big[(\delta+\gamma)^{2\tau} - \delta^{2\tau}\Big] \;=\; 2\gamma \sum_{\tau=0}^{T-1} \tau\, \delta^{2\tau-1} + O(\gamma^2).$$

Consequently, for any two nodes i and j with $|P_i(t;T) - P_j(t;T)| > |R_i(t;T)| + |R_j(t;T)|$, the ordering by trajectory energy coincides with the ordering by own-persistence.

Proof. Expand $\Phi(t,\tau) = \prod_s (D_{t+s} + C_{t+s})$ into 2^τ words. The all-D word applied to e_j is $p_j(t,\tau)\, e_j$, giving P_j. Every other word contains $k \ge 1$ factors C and has operator norm at most $\delta^{\tau-k}\gamma^k$; there are $\binom{\tau}{k}$ such words, so $\|\Phi(t,\tau)e_j - p_j(t,\tau)e_j\| \le (\delta+\gamma)^\tau - \delta^\tau$. Writing $\|a+b\|^2 \le \|a\|^2 + 2\|a\|\|b\| + \|b\|^2$ with $a = p_j e_j$ (so $\|a\| \le \delta^\tau$) and summing over τ gives $|R_j| \le \sum_\tau \big[2\delta^\tau\big((\delta+\gamma)^\tau - \delta^\tau\big) + \big((\delta+\gamma)^\tau - \delta^\tau\big)^2\big] = \sum_\tau \big[(\delta+\gamma)^{2\tau} - \delta^{2\tau}\big]$. The first-order form follows by differentiating in γ at 0; the ordering statement is immediate. □

Remark (asymmetry at first order). For $\tau = 1$, $\|J_t e_j\|^2 = (D_t)_{jj}^2 + \sum_{i\neq j} (C_t)_{ij}^2$: coupling enters E_j only quadratically while own-persistence enters at zeroth order, and the gap compounds geometrically with the horizon through $\delta^{2\tau}$. The effect is intrinsic to ordered-product measures rather than to the $\tau = 0$ convention, which contributes a constant common to all nodes (empirically, the ablation is null; Sec. 4.3). The implementation retains the perturbation's shifted copies in the K-lag companion blocks, adding terms of identical structure, so the decomposition and bound apply term-by-term.

The regime is thus governed by γ/δ; the coupling-to-persistence ratio introduced in Sec. 3.2 is its fitted-model proxy and serves as the diagnostic of when importance rankings are susceptible to persistence dominance.

3 THE AUDIT

3.1 EMBEDDED GROUND-TRUTH PROBE

The obstacle to auditing node importance on real panels is that no node's true multi-step influence is known. The probe introduces a small region of known ground truth into the analyst's own estimation problem. A synthetic subsystem with exactly computable trajectory-level importance is appended to the real panel, and the augmented panel is standardized and fit using the analyst's own estimator. The fitted importance of the synthetic nodes can then be compared directly with their known values. Unlike a standalone simulation, the probe is estimated alongside the real channels and therefore faces the same finite sample, autocorrelation, cross-sectional dependence, and model capacity constraints as the substantive analysis.

The subsystem (Fig. 1) contains six channels per unit: a driver d (AR(1), $\phi = 0.6$); a two-hop relay chain $d \to a_1 \to a_2$ with coupling $c = 0.9$; a control target b_1 that receives the same push from d but propagates to no other node; an isolated AR(1) node b_2; and an observed regime channel r, generated as a persistent ± 1 Markov chain.

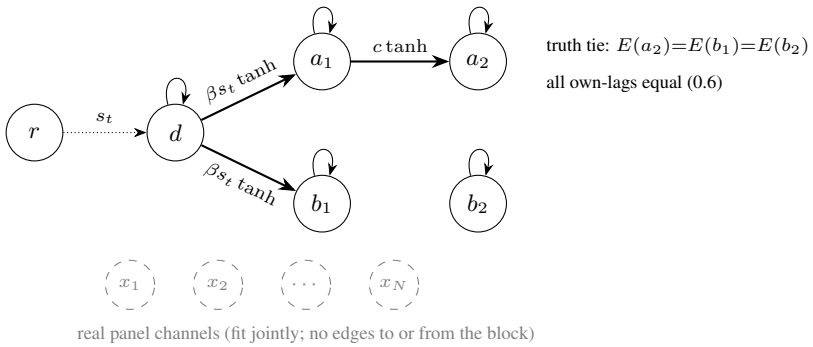

Figure 1: Embedded ground-truth probe. A closed six-channel synthetic subsystem is appended to the real panel and fit jointly with it. Equal own-lags remove persistence as a confound within the probe, while the exactly tied triple $\{a_2, b_1, b_2\}$ provides a per-fit noise floor. The STABLE and REVERSAL arms set $s_t \equiv 1$ and $s_t = r_t$, respectively, distinguishing dynamics that the estimator can represent from those that require interactions outside its model class. True energies are computed from the probe's analytic Jacobians using the same gauge, conventions, and code path as the fitted measure.

Three design choices make the probe diagnostic. First, *all synthetic own-lags are equal* (0.6). Persistence therefore cannot generate importance differences among the synthetic nodes, excluding by construction the confound identified in Proposition 1. Second, a_2, b_1, and b_2 have *exactly equal* true energy. Their fitted spread therefore provides a per-fit estimate of the noise floor against which detected contrasts can be judged. Third, the push from d is generated under two arms. In the STABLE arm, it takes the form $\beta \tanh(d)$, which is representable by an additive function class. In the REVERSAL arm, it takes the form $\beta\, r_t \tanh(d)$, a two-input interaction whose time average vanishes. The two arms distinguish failure to recover representable dynamics from failure caused by restrictions of the estimator class. The regime variable r is observed by the estimator, so failure in the REVERSAL arm cannot be attributed to missing information.

Ground-truth importance is computed in the same gauge and with the same code path as fitted importance. Both use standardized coordinates, companion form, and identical window conventions. For the ground truth, the Jacobians are obtained analytically from the generating equations along the realized trajectory. The resulting truth values and required contrasts, including the a_1/b_1 ratio, the three-way tie, and the REVERSAL trajectory-to-pooled gap, are computed and asserted *before any model is fit*. Runs that fail these construction checks are rejected before estimation.

Because the synthetic block is closed by construction, any fitted *outgoing* coupling from probe nodes to real channels is spurious and provides a direct measure of leakage. Mean absolute outgoing slopes are at most 0.006 across arms and estimators. Fitted incoming couplings from real channels can affect the closed-block truth comparison only at second order in the leakage.

3.2 FITTED-MODEL DIAGNOSTIC: COUPLING-TO-PERSISTENCE RATIO

Proposition 1 motivates a simple diagnostic based on the strength of cross-node coupling relative to own-persistence. When coupling is weak relative to persistence, the persistence term can dominate trajectory energy; stronger coupling gives cross-node propagation greater scope to affect the ranking. We summarize this balance using plug-in quantities available from any fitted model. Let $\bar{J}$ denote the time-averaged absolute lag-1 Jacobian, and define $\hat{\delta} = \mathrm{mean}_i\, \bar{J}_{ii}$ and $\hat{\gamma} = \mathrm{mean}_j \sum_{i \neq j} \bar{J}_{ij}$. We call $\hat{\gamma}/\hat{\delta}$ the *level ratio*. An analogous ratio computed from the dispersion of cross-node and own-node Jacobian terms defines the *spread ratio*. These are pragmatic proxies for the γ/δ of Proposition 1 — means of absolute Jacobian mass rather than the proposition's worst-case norms — chosen for stability on fitted models.

These ratios are diagnostics rather than guarantees of validity. They require no ground truth, refitting, or computation of the importance measure itself, and can therefore be evaluated before running the full audit. Section 5 examines how well they distinguish the observed regimes across five real panels.

4 EMPIRICAL AUDIT OF FITTED IMPORTANCE

We audit the pipeline of an existing research program: a 16-component democracy panel (89 countries, fit window 1950–1995, 4,094 country-years), an additive neural Granger VAR with per-input subnetworks and ℓ_1-regularization, and 20-seed ensembles of the trajectory measure. The trajectory-leverage results we audit are the program's committed internal artifacts, included in pinned form in the supplementary material. Before introducing the probe, we reproduce the committed pipeline. Validation loss is 0.04294 ± 0.00032, matching the committed value to the reported digit, and the reproduced ensemble weights agree closely with the committed ranking ($\rho = 0.947$). All experiments reproduce bit-exactly across sessions (App. B).

4.1 VALIDATION UNDER REPRESENTABLE DYNAMICS

Table 1: Ground-truth and fitted probe results for the additive estimator in the STABLE arm (8 seeds). The fitted trajectory measure recovers the correct ordering in all seeds at both doses and detects the designed multi-step contrast between a_1 and b_1, which the one-step comparator cannot recover by construction. All values are re-derived by the verifier from `hstep/results_hstep.csv`.

	$\beta = 0.5$	$\beta = 1.0$
calibration: d / a_1 / tie	0.71 / 0.81 / 0.91	0.69 / 0.75 / 0.90
a_1/b_1 fitted (true)	1.47 (1.63)	1.40 (1.64)
tie-node spread (CV)	0.01	0.02
ρ(order), d top, a_1 second	0.894, 8/8, 8/8	0.894, 8/8, 8/8
granger-from-fit a_1 / b_1	0.044 / 0.026	0.041 / 0.025

Table 1 reports the STABLE arm, whose generating dynamics lie within the additive estimator class. The fitted trajectory measure recovers the correct ordering of the five truth-known nodes in all 8 seeds at both doses, with d ranked first and a_1 second ($\rho(\text{order}) = 0.894$ reflects only the arbitrary ranking within the exactly tied triple, which has no true order). It also recovers the designed multi-step contrast. Nodes a_1 and b_1 receive the same one-step push, so the one-step comparator cannot distinguish their downstream roles. At $\beta = 0.5$, the true trajectory-energy ratio is $a_1/b_1 = 1.63$ and the fitted measure yields 1.47; at $\beta = 1.0$, 1.40 against a true 1.64. At both doses the detected contrast (the fitted ratio's excess over the exact-tie value of one) is more than 20 times the tie-node spread.

Absolute calibration is less exact and declines with the amount of propagation involved. It is highest for the pure-persistence tie nodes (0.90–0.91), lower for the one-hop node a_1 (0.75–0.81), and lowest for the driver d (0.69–0.71), a pattern consistent with mild shrinkage of learned couplings. The STABLE arm therefore provides a positive control: when the relevant dynamics are representable by the estimator, the fitted trajectory measure recovers the designed ordering and the multi-step distinction that the one-step comparator misses.

4.2 FAILURE MODE 1: ESTIMATOR CLOSURE IS INHERITED BY THE MEASURE

In the REVERSAL arm, the push from d enters through the interaction $r_t \cdot \tanh(d)$. The additive estimator cannot represent this two-input interaction, and because the push reverses sign with r_t, its additive approximation is approximately zero. Table 2 shows the resulting failure and the effect of changing the estimator class. Under the committed additive estimator, the fitted one-step slope from d to a_1 is 0.006, compared with a true value of 0.479. The interaction is therefore effectively absent from the fitted dynamics rather than merely attenuated. The driver's fitted energy falls to its interaction-free level, with calibration of 0.285, and d ranks last among the truth-known nodes and 20th among all 22 channels. We then replace only the additive estimator with a joint per-target MLP, leaving the panel, probe, ground truth, and importance measure unchanged. The joint estimator recovers the slope to 0.347 and the driver calibration to 0.719. It recovers the correct probe ordering in 7 of 8 seeds and raises the driver's mean rank to 4.1 of 22.

The estimator swap localizes the failure. The downstream importance measure is unchanged, but its output changes sharply when the fitted model can represent the interaction on which the true

Table 2: Ground-truth and fitted probe results in the REVERSAL arm ($\beta = 1$) under two estimator classes. The additive estimator cannot represent the interaction $r_t \cdot \tanh(d)$, whereas the joint MLP can. All values are re-derived by the verifier from `results_hstep.csv` and `results_hstep_joint.csv`.

	additive	joint	truth
one-step slope $d \to a_1$	0.006	0.347	0.479
calibration of d	0.285	0.719	1
ρ(order) / d top	0.447 / 0/8	0.880 / 7/8	1 / –
rank of d, all 22 channels	20.0	4.1	–

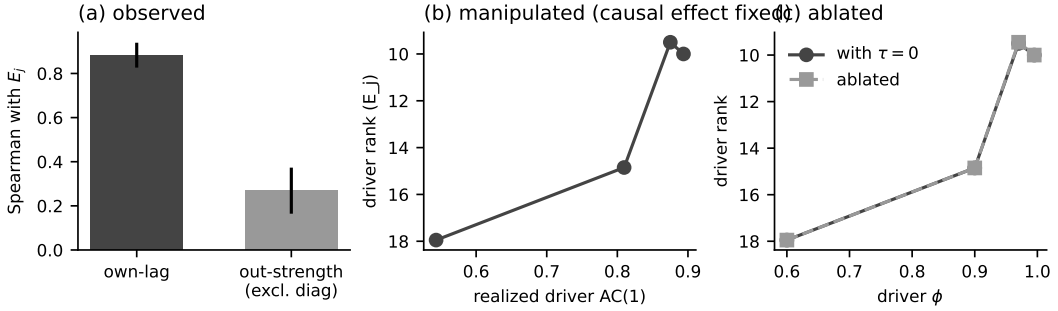

Figure 2: Persistence weighting on the audited panel. (a) Observed: the trajectory-importance ranking is strongly associated with fitted own-lag mass but only weakly with diagonal-excluded outgoing influence (20 seeds, mean $\pm$ sd). (b) Manipulated: increasing only the injected driver's persistence, while holding its causal effect fixed, improves its rank and raises its share. The non-monotone tail occurs where near-unit-root persistence becomes difficult to estimate on 46-year segments. (c) Ablated: removing the $\tau = 0$ self-term leaves the persistence effect essentially unchanged, showing that it arises from multi-step propagation rather than from inclusion of the initial perturbation.

importance depends. Fitted importance therefore inherits the representational limits of the estimator from which it is computed. The richer estimator also introduces costs relevant to interpretation. Its coupling recovery is asymmetric across seeds despite symmetric ground truth, indicating an identifiability problem (Anonymous, 2026a). Moreover, its predictive error is slightly lower than that of the additive estimator (0.036 vs. 0.038 MSE), yet on the unmodified panel it reverses the substantive ordering of the real channels (rank correlation -0.50 with the committed ordering; App. B). Similar predictive accuracy can therefore coexist with substantially different importance rankings.

4.3 FAILURE MODE 2: PERSISTENCE WEIGHTING OF TRAJECTORY IMPORTANCE

On the real channels, the audit finds the persistence-dominated regime motivated by Proposition 1. Three tests provide complementary evidence (Fig. 2). *Observed:* across 20 baseline seeds, the trajectory-importance ranking correlates strongly with fitted own-lag mass ($\rho = 0.883 \pm 0.056$) but much less with diagonal-excluded outgoing influence ($\rho = 0.269 \pm 0.104$). Partial Spearman correlations, controlling each predictor for the other, preserve the contrast at $+0.864$ for own-lag mass and $+0.254$ for outgoing influence. *Manipulated:* an injected driver with a fixed, calibrated causal effect remains near the bottom of the trajectory-importance ranking across a fourfold coupling-dose range, with mean rank 17.95 of 18 and recovery-at-3 equal to 0.00 at every dose. The largest dose matches the largest fitted coupling among the real channels. Holding this causal effect fixed and increasing only the driver's own persistence changes the result substantially: its mean rank improves to 9.5 from 17.95 and its importance share nearly triples. Thus, the fitted importance changes substantially when persistence changes even though the injected cross-node effect does not. *Ablated:* removing the $\tau = 0$ self-term changes the observed rank ranges by at most 0.05. The persistence weighting therefore does not arise from simply counting the initial perturbation in the energy. It persists through the subsequent ordered Jacobian products, consistent with the persistence contribution isolated in Proposition 1. The upper end of the persistence sweep also reveals an estimation bound-

Table 3: Five panels; the coupling-to-persistence ratio orders persistence dominance; time scale does not. `crosspanel/crosspanel_seed_stats.csv`.

panel	own-lag mean	level ratio	$\rho(E_j, \text{own})$	$\rho(E_j, \text{out})$	partial out|own
democracy	0.855	0.63	0.86	0.21	+0.30
wdi	0.641	1.24	0.80	0.12	-0.49
realized vol	0.295	1.26	0.85	0.31	+0.62
air quality	0.640	1.73	0.51	0.09	+0.11
macro	0.694	2.58	0.55	0.79	+0.70

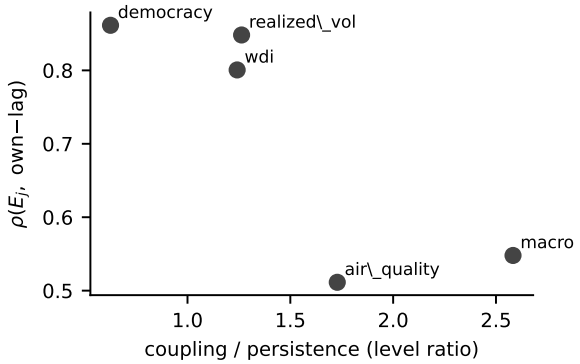

Figure 3: The coupling-to-persistence level ratio orders persistence dominance across five panels; time scale does not (realized volatility is the fastest panel and among the most persistence-dominated).

ary. The response becomes non-monotone as the injected process approaches unit-root persistence. On 46-year segments, persistence in this regime is poorly identified, and recovery peaks where the injected persistence is estimated most accurately rather than where its true value is largest.

4.4 Failure mode 3: pooled linearization off the data manifold

Static reductions replace trajectory-dependent Jacobians with a single Jacobian evaluated at a reference state, typically the sample mean. The probe reveals a distinct failure mode in this reduction. In the REVERSAL arm, the mean state has $r \approx 0$, even though the regime channel takes only the values -1 and $+1$; the pooled linearization is therefore evaluated at a state that is never observed in the data. The two estimator classes behave differently at this unsupported state. The additive model, which cannot represent the interaction, underestimates the pooled energy (3.5 vs. a true value of 4.7). The joint model instead extrapolates its learned interaction to $r \approx 0$ and overestimates the pooled energy by a factor of 2.6 (12.4 vs. 4.7). Consequently, the trajectory-to-pooled gap, the quantity used by departure-style diagnostics, is inaccurate under both estimator classes, but for different reasons (1.26 and 0.92, respectively, compared with a true value of 3.25). The same issue appears even in the STABLE arm. There, fitted pooled energies have cell means of 7.7–9.7, compared with true values of 21.6–40.8. A point linearization evaluated in a region with little or no data support therefore reflects the model's extrapolation behavior as well as the underlying dynamics. Trajectory-averaged reductions do not exhibit this failure in our experiments. More generally, importance statistics based on derivatives evaluated away from the observed trajectory require explicit checks that the reference state is supported by the data.

5 Scope: Conditions for Each Failure Mode

Table 3 and Fig. 3 place five real panels on the diagnostic. Two patterns are immediate. First, persistence dominance is not simply a time-scale phenomenon: the fastest panel (realized volatility, mean own-lag 0.295) is among the most persistence-ordered ($\rho = 0.85$), while the slower democracy and development panels show the same pattern. Second, the coupling-to-persistence level ratio

orders the phenomenon: persistence dominance weakens as off-diagonal propagation approaches twice the own-lag mass, consistent with Proposition 1. With five panels, this is an ordering rather than an estimated response surface.

The macroeconomic panel appears to be an exception: the measure tracks outgoing influence ($\rho = 0.79$). The probe shows why this association is not sufficient to establish sensitivity to influence. Persistent nodes on this panel are also influential ($\rho(\text{own}, \text{out}) = 0.735$), so the two explanations are observationally confounded. Injecting a driver whose influence and persistence are controlled independently resolves the confound. Across a fourfold range of calibrated doses, reaching about three-quarters of the injected link's saturation at the top, the trajectory measure never places the driver in the top three (recovery-at-3 0.00 across mechanisms and doses), whereas the one-step out-strength comparator recovers it at mean rank 2 from the middle dose upward. The model learns the injected effect, but the horizon-propagated measure does not surface it. A dose beyond the valid range (training error below the unaugmented baseline) is reported as a boundary condition in App. D. The remaining pattern, strong coupling with homogeneous persistence in the air-quality panel, yields rankings that track neither statistic stably; we report this case as unidentified rather than as a fourth regime. The practical scope is therefore conditional in both directions. A low ratio indicates susceptibility to persistence-dominated rankings; a high ratio removes that warning but does not establish sensitivity to influence. The probe, rather than the diagnostic alone, remains the audit's primary validation instrument.

6 Case Study: Auditing a Research Program's Committed Ordering

The audit was motivated by a research program that applies this measure–estimator composition to a 16-component democracy panel (89 countries, 1950–1995 fit window). Its committed 20-seed ensemble ranks Freedom of Association and Freedom of Expression as the highest-leverage components and Suffrage as the lowest. We reproduce the pipeline exactly (Sec. 4) and ask what drives this ordering. The committed weights correlate strongly with fitted own-lag mass ($\rho = 0.944$) but weakly with diagonal-excluded outgoing influence ($\rho = 0.126$). Partial Spearman correlations preserve the contrast: $+0.864$ for persistence given influence and $+0.254$ for influence given persistence. Individual inversions illustrate the difference. Suffrage ranks 16th of 16 on the committed measure but 3rd on outgoing influence, while Clean Elections ranks 3rd on the committed measure but 14th on influence (Fig. 4).

The pipeline's committed corroboration statistic, an alignment between the measure and "out-strength" ($\rho = 0.803$) recorded in its internal artifacts, does not contradict this result. The committed out-strength definition sums full columns of the mean Jacobian, including the diagonal, and therefore incorporates own-persistence. Removing the diagonal reduces its correlation with the committed weights to 0.247. These findings do not overturn the pipeline's dissociation of the measure from variance or its predictive results. They do change the interpretation of cross-component leverage. The democracy panel has the lowest coupling-to-persistence ratio among the five panels (0.63), placing it in the persistence-dominated regime identified by the audit. Cross-component importance differences on this panel therefore cannot be interpreted as differences in outgoing influence without additional validation. Comparisons within persistence strata or calibration with an embedded probe provide two ways to separate the two quantities. This is a conclusion the original analyses had no instrument to reach; providing that instrument is the audit's purpose.

7 Related Work

Our audit connects four literatures. *Network controllability* provides the importance measures we test, including Gramian-based and modal controllability metrics (Pasqualetti et al., 2014; Liu et al., 2011), widely applied in network science (Gu et al., 2015). These metrics are typically defined for known dynamics; we ask what they measure when the dynamics must first be estimated. *Neural Granger and structured neural VAR models* provide the estimator classes we audit, including additive per-input architectures (Bussmann et al., 2021) and joint sparse-input networks (Tank et al., 2022); recent applications of this model family to institutional and democracy panels (Zaytsev et al., 2026; Kuskova et al., 2026) illustrate the interpretive practices the audit is designed to check. Sec-

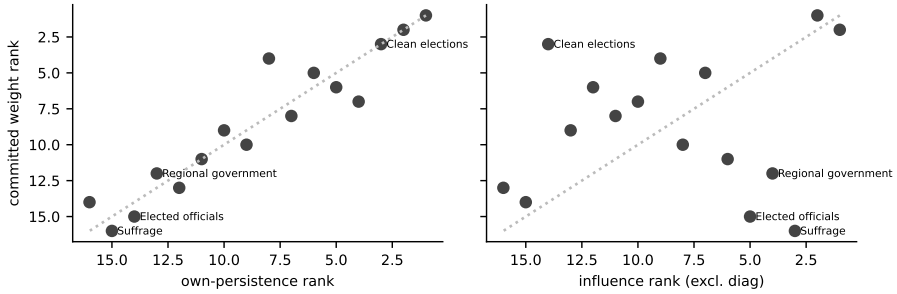

Figure 4: The committed importance ordering tracks fitted own-persistence (left) much more closely than diagonal-excluded outgoing influence (right). Labeled nodes show the inversions discussed in the text.

tion 4.2 shows how restrictions of the estimator class propagate into downstream importance even when predictive fit remains similar; companion work studies identification of interactions within these architectures (Anonymous, 2026a). *Attribution evaluation* provides the closest methodological precedent. Model- and label-randomization tests (Adebayo et al., 2018) and remove-and-retrain benchmarks (Hooker et al., 2019) show why explanations require constructed validation tests rather than visual inspection; integrated gradients (Sundararajan et al., 2017) is one of our comparators. We extend this logic to dynamical importance, where ground truth is a multi-step, gauge-consistent quantity embedded in the analyst's own data and estimated jointly with it. This approach also follows calls for empirically grounded evaluation of causal models (Gentzel et al., 2019). Finally, the trajectory-importance measure instantiated here is developed in companion work (Anonymous, 2026b), while classical VAR theory provides the companion-form machinery (Lütkepohl, 2005).

8 Practice, Limitations, and Conclusion

Practice. The audit adds hours rather than weeks to an analysis. (1) Compute the coupling-to-persistence ratio from the fitted mean absolute lag-1 Jacobian. A low ratio flags susceptibility to persistence-dominated rankings; substantive cross-node comparisons then require additional validation or comparison within persistence strata. (2) Run the embedded probe under the analyst's own panel, estimator, gauge, and measure code path. Inspect calibration on the truth-known nodes, the multi-step contrast relative to the tie-node noise floor, and, where interactions are plausible, the gap between the representable and interaction arms. (3) Check data support for any reference state at which fitted derivatives are evaluated; pooled linearizations failed in opposite directions under our two estimator classes. (4) Do not treat agreement between diagonal-inclusive statistics as independent convergent validity.

Limitations. Exact ground truth applies only to the embedded block; the importance of real channels remains unknown, and the probe validates an estimator-measure composition under designed conditions rather than any substantive claim. Our scope results cover five panels and two estimator classes, so the ratio diagnostic establishes a cross-panel ordering, not an estimated response surface or universal threshold. The probe's regime channel is observed by construction, isolating representational closure but not addressing latent regimes. Finally, we study one trajectory-measure family. The persistence mechanism follows from ordered Jacobian products (Proposition 1), while representational closure can affect any downstream measure computed from fitted derivatives, but their empirical generality beyond the settings studied here remains to be tested.

Conclusion. Importance measures derived from learned dynamical models can be tested against ground truth without abandoning the analyst's empirical setting. The embedded probe provides a positive control when the estimator represents the relevant dynamics and exposes distinct failures due to representational closure, persistence weighting, and unsupported linearization. The audit is inexpensive and turns validation of fitted importance from inspection into an empirical test.

Reproducibility statement

All experiments run in a pinned pipeline whose design document and study library are SHA-gated before any fit; every deviation is logged (App. B). The grid ($\sim$800 fits) is deterministic: a deliberate re-run of a complete 20-fit cell reproduced every stored score to one unit in the last place (maximum absolute difference 2^{-53} over 2,880 values; identical GPU class), and an accidental same-session duplicate of a separate cell, compared symmetrically through serialization, was exactly bit-identical (run log). Every number in this paper is recomputed from the released artifacts by an automated verifier (a tolerance-typed claim registry whose report ships with the artifacts); the manuscript convention is that no value appears outside the registry. Code, artifacts, design document, and verifier are provided in the supplementary material for the review.

AI use statement

Generative AI was used in this paper in three limited roles: copy-editing author-written prose for clarity, supporting the development and debugging of analysis code, and formatting references. All substantive content, including methodology, analyses, and results, were created and executed by the authors. Every AI-assisted edit was compared against the original to confirm that meaning and claims were preserved; all AI-assisted code was reviewed, tested, and confirmed by the authors to reproduce the reported results; and formatted references were verified against their sources. The authors accept full responsibility for the final work.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems 31*, pp. 9525–9536, 2018.

Anonymous. Companion submission on identifiability of neural interaction discovery (under review; cited anonymously), 2026a.

Anonymous. Companion submission introducing the trajectory-leverage measure (under review; cited anonymously for double-blind compliance), 2026b.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science (DS 2021), Lecture Notes in Computer Science vol. 12986*, pp. 446–460. Springer, 2021.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems 32*, 2019.

Shi Gu, Fabio Pasqualetti, Matthew Cieslak, Qawi K. Telesford, Alfred B. Yu, Ari E. Kahn, John D. Medaglia, Jean M. Vettel, Michael B. Miller, Scott T. Grafton, and Danielle S. Bassett. Controllability of structural brain networks. *Nature Communications*, 6:8414, 2015. doi: 10.1038/ncomms9414.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems 32*, pp. 9737–9748, 2019.

Valentina V. Kuskova, Dmitry Zaytsev, and Michael Coppedge. Function-valued causal influence in nonlinear time series. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. Controllability of complex networks. *Nature*, 473(7346):167–173, 2011.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

Fabio Pasqualetti, Sandro Zampieri, and Francesco Bullo. Controllability metrics, limitations and algorithms for complex networks. *IEEE Transactions on Control of Network Systems*, 1(1):40–52, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

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

Dmitry Zaytsev, Valentina V. Kuskova, and Michael Coppedge. From causal discovery to dynamic causal inference in neural time series. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26)*, Jeju Island, Republic of Korea, 2026. ACM. doi: 10.1145/3770855.3818956.

A PROBE CONSTRUCTION AND EXACT TRUTH

Generating equations. Per unit, with burn-in 50 steps discarded and $s_t \equiv 1$ (STABLE) or $s_t = r_t$ (REVERSAL):

$$\begin{aligned} d_{t+1} &= \phi_d\, d_t + \eta_t^d, \quad \eta_t^d \sim \mathcal{N}(0, 1-\phi_d^2), \qquad \phi_d = 0.6; \\ a_{1,t+1} &= \rho\, a_{1,t} + \beta\, s_t \tanh(d_t) + \varepsilon_t, \qquad b_{1,t+1} = \rho\, b_{1,t} + \beta\, s_t \tanh(d_t) + \varepsilon_t'; \\ a_{2,t+1} &= \rho\, a_{2,t} + c\, \tanh(a_{1,t}) + \varepsilon_t'', \qquad b_{2,t+1} = \rho\, b_{2,t} + \varepsilon_t''', \end{aligned}$$

with $\rho = 0.6$ for every synthetic own-lag, $c = 0.9$, innovation sd 0.5, and r_t a ± 1 Markov chain with stay probability 0.95. Driver, regime, and relay noise use disjoint, mechanism-keyed RNG streams so that arms and doses are counterfactually comparable.

Exact truth in the fitted gauge. With σ the fit-window standard deviations of the synthetic channels and $D_\sigma = \mathrm{diag}(\sigma)$, the true standardized one-step Jacobian is $J_t^{\mathrm{std}} = D_\sigma^{-1} J_t^{\mathrm{raw}} D_\sigma$, with J_t^{raw} read off the equations above (sech^2 terms at the realized states). True energies embed J_t^{std} in the same $K{=}3$ companion form, with the same window starts, the same $\tau{=}0$ convention, and the same energy code path as the fitted measure; pooled truth linearizes at the fit-window mean state (with the realized regime mean), realized truth at the time-averaged J^{std}. Because the block is closed, these truths are exact for the full system.

Pre-fit verification. Before any training run, the construction is asserted on the generated data: the a_1/b_1 truth ratio exceeds 1.4 (realized 1.63–1.70 across arms and doses); the energies of a_2, b_1, b_2 tie exactly; the one-step inputs $d \to a_1$ and $d \to b_1$ agree within sample-sd tolerance; and in the REVERSAL arm the trajectory-vs-pooled truth gap for d exceeds 2. A construction failing any assertion never consumes GPU time (see App. B).

B DESIGN DOCUMENT, DEVIATIONS, AND CAUGHT ERRORS

The audit's credibility rests on the standard it applies, so we record what its own pipeline caught. (1) *Estimator identification.* The program's internal handoff described the democracy estimator as a joint neural VAR; reproduction of the committed ensemble failed (rank correlation -0.50; regenerated as a pinned artifact) until the estimator was identified from the committed notebook as the additive class, after which reproduction was exact. Audit pipelines must pin the estimator by artifact, not description. (2) *Void manipulation.* The first persistence sweep ran entirely at $\phi = 0.6$: the sweep mutated a module constant that a generator had bound as a default argument at import. Cost, four GPU-hours; remedy, the pre-fit verification of App. A. (3) *Silent data-coding hazard.* The condition label `NULL` is parsed as missing by default CSV readers; one resume pass re-ran a completed cell and appended a duplicate. The duplicate was shown bit-exact against the original (maximum absolute score difference 0; logged) and removed; all readers were pinned. Prompted by this accident, determinism was then demonstrated deliberately: a full re-run of a complete 20-fit cell reproduced every stored score to one unit in the last place of double precision (maximum absolute difference 2^{-53} over 2,880 recomputed values, identical GPU class; the residual is serialization round-trip, not pipeline nondeterminism), shipped as a pinned artifact. (4) *Dose validity.* On one panel the

pre-registered calibration rule was unsatisfiable (the injected link saturates below the largest fitted coupling); the boundary run produced training error *below* baseline — targets dominated by injected signal — and was reclassified as a boundary condition (App. D), with doses re-set on the slope scale. The SHA-pinned design document and a run log recording each deviation accompany the artifacts.

C COMPARATOR PATHOLOGIES

Modal controllability. Fails the null-panel sanity check: across 20 seeds, disconnected AR(1) decoys exceed the median real channel in 80–100% of seeds; on the macro panel modal ranks the injected driver first, consistent with a preference for fast modes — an anti-persistence pathology, not sensitivity to influence.

Within-unit variance. Non-comparable by construction on appended-channel designs: injected channels carry no between-unit level structure, so their within-unit variance is inflated relative to real channels whose variance is largely between-unit. All five disconnected decoys rank 1–5 of 22 on this statistic, raw and standardized alike.

Integrated gradients through rollouts. In the linear limit the IG-through-time score of channel d reduces to $(G\Sigma_x)_{dd}$ — a data-covariance-weighted diagonal, not a Gramian trace: trajectory attribution and trajectory leverage differ before any nonlinearity, which is why IG's dose sensitivity coexists with bottom ranks in the grid.

Null-cell calibration. Disconnected decoys' renormalized share is 0.0194 (95th percentile 0.0205); the *persistent* no-effect regime channel sits mid-table (share 0.054) — the persistence result reproduced inside a cell where nothing causal occurs.

D ADDITIONAL TABLES

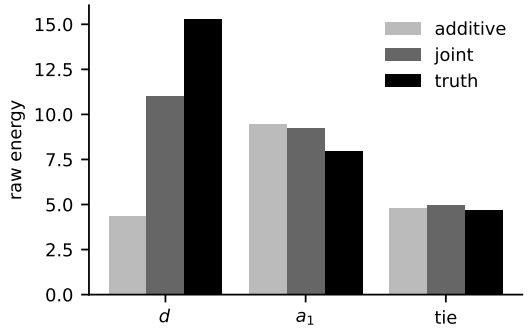

Figure 5: Probe energies under the two estimator classes against exact truth (REVERSAL arm, $\beta = 1$): the additive class misses the driver's interaction content entirely; the joint class recovers most of it and slightly overshoots the one-hop relay.

Table 4: Dose insensitivity on the audited panel (driver's renormalized E_j share, rank in parentheses; $p = 0.95$, primary injection seed, 20 fits per cell). NULL reference: share 0.0194, 95th percentile 0.0205.

mechanism	β_{lo}	β_{mid}	β_{hi}
sign-stable	0.0196 (17.95)	0.0198 (17.95)	0.0202 (17.95)
sign-reversal	0.0196 (17.95)	0.0198 (17.95)	0.0202 (17.95)
reroute	0.0197 (17.95)	0.0201 (17.95)	0.0207 (17.95)
six-target	—	—	0.0213 (17.95)

Table 5: Macro overdose boundary condition (doses from the saturating calibration rule; the record of why dose-validity gates matter). Real-channel training MSE falls *below* the unaugmented baseline (cell means 0.25–0.30 vs 0.37) while a one-step comparator detects the driver at rank 1–2 and the trajectory measure leaves it at rank 7.4–10.6 of 11.

	granger rank	E_j rank
reroute, β=10/20	1.0 / 1.0	7.4 / 10.0
sign-reversal, β=10/20	2.0 / 2.1	7.9 / 8.0

Replicated injection seeds (four per critical cell, 8 fits each), the six-target variant, and complete per-seed probe results are in the released artifacts; per-fit diagnostic flags (fit quality, learnability) are recorded throughout.