DEPSHIFT: MEASURING WHAT CROSS-CHANNEL MODELING BUYS, AND WHAT IT COSTS UNDER DEPENDENCY SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

A recent method proposes channel shuffling as a certificate that a multivariate forecaster does not over-rely on cross-channel structure. It certifies less than it is taken to. Shuffling is a channel *permutation*, a relabeling of a coupling that stays intact; the disturbance a deployed model faces is a *rewiring*, the coupling changing with channel identities fixed. A relabeling also permutes the per-channel marginals unless the channels it swaps share one law, whereas a change in the coupling's lag leaves every single-channel law exactly unchanged. We build *DepShift*, a generator on which the lag change is the pure cross-channel disturbance and a relabeling is not, and an *excess-fragility* contrast against a channel-independent control. Permutation-equivariance does not confer robustness to rewiring: the same model is more fragile to a rewiring than to a relabeling at 11 of 12 coupling configurations, null in the factor topology, and the flip holds for CPiRi's released code at 5 of 5, uncorrected by its own shuffling regulariser. Along a dose axis, and replicated on six fresh seeds, the equivariant model's response to a lag change grows far faster than its response to a relabeling. What permutation degradation it does show is its evaluation pipeline's, and the same normalisation switch leaves the fragility of all three identity-storing architectures we put through it in place. The two classes also separate out of sample at full channel count. To read a null one must know what was achievable, so we compute a closed-form *ceiling* for the cross-channel increment: against it the equivariant model pays a tenth to a quarter of the optimal predictor's price under a lag change and under a twentieth under relabeling. We report where this fails: the ceiling does not transfer to real data, 3 pre-registered attempts to attribute the real-data ordering to coupling movement failed, the real-data test is met on one of three datasets, and a ratio-metric artifact produced a retired hypothesis's predicted sign at $p{=}2e-59$ inside a protocol written by authors trying not to be fooled.

1 INTRODUCTION

Channel-independent models match or beat channel-dependent ones on standard benchmarks (Nie et al., 2023; Zeng et al., 2023), prompting the view that channel dependence buys capacity at the price of robustness (Han et al., 2024). That framing bundles whether cross-channel information is *useful*, largely a property of the *data*, with what modeling it *costs* under shift, partly a property of the *architecture*. Benchmarks factorize shift along several axes (Wiles et al., 2022) but not inter-channel coupling, so "CI is robust" is a correlational observation over datasets of unknown coupling, and the shift literature attends to *marginal* drift (RevIN, Kim et al., 2022), and *dependency* shift — the inter-channel structure itself changing — is rarely isolated as an axis of its own; §2 places the work that comes closest.

For the architectural half one recent method proposes a diagnostic: shuffle the channels and see what breaks, which Xu et al. (2026) reports and builds a method around. How widely it is relied on we do not measure (Appendix N). Our main result is that this certifies less than it is taken to. Shuffling is a *permutation*, a relabeling of a coupling that is otherwise intact, while the disturbance a deployed model faces is a *rewiring*, the coupling itself changing with channel identities fixed. Permutation-equivariance covers only the first, and a model certified robust by shuffling is measur-

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.

Table 1: What this paper establishes, and what it does not; bold rows are claims we set out to make and could not.

Claim	Evidence	Status
Equivariance to relabeling is not robustness to rewiring	11/12 coupling configs, sign test p=0.0032 descriptive (configs share seeds; null in the second topology); CPiRi itself 5/5 configs, nominal p=0.031; at matched disturbance size $+14.3$ vs -0.2%p; dose interaction replicated on six fresh seeds (E34), $+40.2$%p, p=0.0312	supported
Channel-identity storage *predicts* permutation fragility	5 prospective arms of 7, perfect separation, exact p=0.10	suggestive
. . . and survives adaptive normalisation	E31: two of three arms by both registered criteria, the third by one; the equivariant arm's vanishes	supported, per arm
. . . and *causes* it	pre-registered intervention: removal moves the wrong way	**not shown**
The ordering holds out of sample	full C: Traffic separates all 3 identity-storing arms (15.2%p vs 1.4%p floor); Electricity 6.3%p on one, interval includes zero at the window unit	supported
. . . attributable to coupling movement	three pre-registered designs, the third separating cross-channel from marginal movement; nothing	**failed three times**
Rewiring can be run with marginals held fixed	exactly, by the lag change (Lemma 1); not by a relabeling, which permutes per-channel variance (0.19 at k=8) and cannot be rescaled onto the manifold in 10 of twelve cells (E32)	supported
Cross-channel value is scarce but not absent	closed-form ceiling: directed-lag $+31$%, every other family < 6%	supported
. . . the ceiling transfers to real data	three principled repairs; the last flips sign with the split	**not shown**
Recoverability moderates cross-channel value	three generalisation tests; available value moves the other way	**retired**

ably more fragile to the second (§6). Nor are they one disturbance at two strengths: a relabeling of channels with different laws moves the per-channel marginals and a lag change moves none of them (Lemma 1), so what an equivariant model shows under shuffling is its evaluation pipeline, while the identity-storing models we test keep their fragility once that pipeline is fixed (§6). Every claim above was pre-registered with a kill criterion; Table 1 lists 10, of which 5 are not established.

Establishing that required an instrument, and building one exposed a third obstacle: comparing a trained CD with a trained CI model says something about two models, and about the data only if one knows what was achievable. Estimating that ceiling is routine elsewhere (Granger, 1969; Geweke, 1982; Williamson et al., 2023) but not on the *channel* axis, and an earlier version called cross-channel information *redundant* where the ceiling shows a gain was available and our models failed to take it (§5). So we contribute instruments and report what they measure, including where our hypotheses failed.

Contributions. **(1)** An architectural boundary (§6). On a dose axis, in a pre-registered replication on six fresh seeds with the seed as the unit, an equivariant model's lag-change slope exceeds its relabeling slope by $+40.2$%p per unit dose (p=0.0312; Fig. 1). Its relabeling slope, $+4.4$ $[-9.2, +18.0]$, is small and consistent with the zero its symmetry predicts; we test the difference, not that zero. The ordering holds at 11 of 12 coupling configurations, on real data at full channel count, and on CPiRi itself. A pre-registered audit of channel-identity parameters predicts permutation fragility on the arms it was written for and fails to generalise to six it had not seen; what permutation fragility the equivariant model shows is pipeline, and the audit's *causal* version does not go through. **(2)** DepShift and the excess-fragility contrast (§3): controlled coupling with paired clean/shifted targets, a lag change that holds every single-channel law fixed exactly against a relabeling that cannot, and a CI-control subtraction; an operational contrast, not an identified causal effect. **(3)** A closed-form ceiling for the cross-channel increment (§5), released as a training-free diagnostic; the idea is Granger (1969)'s and Williamson et al. (2023)'s, computed rather than estimated here (§2), and it reclassifies two of our own nulls. **(4)** What the instruments report, including against us (§4, §7).

2 RELATED WORK

Channel independence vs. dependence. Surveyed by Qiu et al. (2026). Han et al. (2024) frame a capacity/robustness trade-off between CI and CD *as a training strategy on a fixed backbone*: for a linear model, CI depends on the *mean* autocorrelation across channels, which drifts less between train and test than most individual channels' do, and they propose a regularizer (PRReg). We differ on both axis and cause: our shift axis is inter-channel *coupling* change (not marginal drift), and our cause is *architectural* (channel-identity parameters), not a training choice.

Inter-series shift, and generators for it. Closest on our shift axis are JointPGM (He et al., 2025), which models intra- and inter-series transitional shift jointly, and Vapsi et al. (2026), whose synthetic series carry time-varying, regime-switching correlations and cross-channel lag structure. Both differ from us in role — a forecaster and a data source, built for or on data whose coupling change is unobserved — where ours is an instrument in which the change is *known* and paired against an otherwise identical clean series, so a degradation can be attributed rather than observed.

Predictability ceilings, and where ours differs. Estimating what is achievable before judging a model is established practice (Feng et al., 2026; Catt, 2026; Tan et al., 2025), and within the CI/CD debate Abdelmalak et al. (2026) use Granger tests to argue CI's advantage reflects weak inter-channel dependence. Our ceiling is not new: it is Granger (1969) and Geweke (1982)'s restricted-versus-unrestricted comparison, and a special case of Williamson et al. (2023)'s oracle variable importance — the group form of Lei et al. (2018)'s leave-one-covariate-out contrast — with the other channels excluded. What we add is that on these generators it is *computed* rather than estimated, which reclassifies individual CI-vs-CD nulls as facts about the data or about the model. Variable-importance permutation (Fisher et al., 2019; Hooker et al., 2021; Chamma et al., 2023) differs from ours: it shuffles one variable's *values* across observations, breaking its link to the target, while a relabeling renames every channel of input and target together and breaks nothing but the index.

Brigato et al. (2026) find that small setup changes reorder the leaderboard across eight long-horizon forecasters and 14 datasets, which is why we subtract a control, pre-register what we test, list where those registrations failed, and make no ranking claim.

Permutation equivariance across channels. The relabel/rewire distinction is established outside this literature: Kenlay et al. (2021) bound how far a graph convolution moves under *degree-preserving* rewiring, the same disturbance on a graph that relag is on a coupling, and Gama et al. (2020) separate permutation symmetry from stability to that disturbance; and Wu et al. (2022) make the graph's structure a first-class part of the out-of-distribution formulation, editing it adversarially to build environments ("rewiring" there also names an intervention (Topping et al., 2022), a sense we do not use). In forecasting, Jeong & Suk (2026) characterise which linear state-space models are channel-equivariant and Jang et al. (2026) take dependency as one of three axes to be robust to; neither varies the coupling at test time. What we transfer is the instrument separating the two axes, to a literature where we counted the inference from shuffling to structural robustness once (Appendix N). Closest here is CPiRi (Xu et al., 2026), which argues that channel-dependent models' permutation fragility is architectural, that they "overfit the channel ordering", and that channel permutation-*invariance* is the cure, enforced by a channel-shuffling objective and diagnosed by shuffling at test time. Its stated scope is what makes it the right target: it offers this as "enabling deployment in settings with *structural* and distributional co-drift without retraining" (emphasis ours), scoping that as "deployments where the channel set or topology changes", and tests channel permutation and channel-set change. What neither its objective nor either diagnostic varies is the coupling *function* with the channel set held fixed, and that is the disturbance we isolate. Our audit therefore puts an attribution CPiRi makes qualitatively into pre-registered controlled form; what we add is that second axis, and §6 measures CPiRi itself on it. The backbones we test span a channel-identity spectrum, from iTransformer (Liu et al., 2024), whose shared variate map has no per-variate parameter, to Crossformer (Zhang & Yan, 2023), TSMixer (Chen et al., 2023) and TimeXer (Wang et al., 2024b), each of which stores channel identity in a parameter.

3 DEPSHIFT AND THE EXCESS-FRAGILITY PROTOCOL

DepShift generates multivariate series with controlled inter-channel coupling and a test-time change of it: three *topologies* — *diffusion* (symmetric neighbour averaging), *directed-lag* (leader→follower with delay τ), *latent-factor* — a coupling strength ρ, and at a break point one of two *shift* modes we deliberately separate: *rewire* permutes the channel assignment (a relabeling of a fixed coupling, the axis channel-shuffling diagnostics test) while *relag* changes $\tau \to \tau'$ (a non-permutation change of the structure). A model can be invariant to the former and sensitive to the latter. Every synthetic sweep and real-data run in the main text uses lookback 96, horizon 24, 18 epochs, short of the grid this literature reports; nothing below speaks to longer horizons.

One thing a relabeling cannot do is hold marginals fixed: for a stationary linear Gaussian process it permutes each channel's whole autocovariance sequence, so the single-channel laws are preserved only if the channels on each orbit share one such sequence, and ours do not (a $1.8\times$ gradient; Lemma 1, Appendix P). A lag change does hold them, exactly: every single-channel law of the chain is τ-invariant and only cross-covariances move. So the coupling change satisfies this paper's definition of rewiring, the relabeling it is compared against does not, and no innovation rescaling repairs that in 10 of twelve cells (E32, Appendix R).

Excess fragility. Subtracting a control's degradation to avoid attributing baseline difficulty to a mechanism is in the spirit of effective robustness (Taori et al., 2020), though ours is a difference in differences against one stated control. A shift can also perturb per-channel marginals, which a channel-independent (CI) model is already sensitive to and which normalization like RevIN only partly removes, so to isolate the *coupling*-specific effect we report $\text{excess} = \Delta_{\text{model}} - \Delta_{\text{CI}}$, where $\Delta = (\text{MSE}_{\text{shift}} - \text{MSE}_{\text{clean}})/\text{MSE}_{\text{clean}}$ and the CI control (PatchTST) can only be moved by the marginal component.

What to run instead of a shuffling test. A criticism only helps if there is something else to do, so we release the protocol as a function rather than as a description. `timecard.depshift`'s `evaluate` takes a model factory and returns its excess under both shift families, holding fixed three things our own code got wrong at least once: candidate and control see the *same* realization at each seed, since excess is paired and unpairing it silently invalidated an earlier version of our flip comparison; normalisation statistics come from the training segment alone and score the clean and shifted segments both, since re-normalising the shifted one hides the marginal movement the control exists to absorb; and no evaluation window straddles the break.

4 WHAT THE INSTRUMENT REPORTS: VALUE IS SCARCE, AND STRUCTURE SETS IT

Cross-channel *value* is scarce, and where it exists it depends on the *structure* of the coupling rather than on how many channels there are. Under *diffusion* (synchronizing) coupling a channel and its neighbours co-move, so its own past already carries the shared signal; only *directed-lag* coupling — a follower delayed from a leader it does not otherwise track — produces a large gain ($+22\%$ at ρ=0.8). Bolting cross-channel modeling onto a strong backbone two ways is net-negative on homogeneous high-channel benchmarks, but those arms stop at 5–8 epochs against the backbone's 12–14, so part of each Δ is optimization rather than mechanism and they belong in Appendix F, not here. §5 supplies the evidence we do lean on, where "was there anything to gain?" has an exact answer.

We had also built a controlled recoverability knob and found cross-channel *value* and dependency-shift *fragility* falling toward zero together along it ($9.6\%/7.4\%$ to $0.7\%/0.3\%$ over 8 seeds), treated it as a hypothesis, and retired it: it survived none of three pre-registered tests, and §5 shows the co-movement was never in the data (Appendix A).

On real data the sign splits, and not by channel count. Toggling TimeMixer's (Wang et al., 2024a) channel-independence flag — same architecture, same budget, five seeds — helps 76% of Weather's channels ($+1.5\%$ of target variance) while the identical toggle *hurts* 89% of Electricity's. So "cross-channel modeling does not pay" is false as a blanket statement and we do not make it. The

toggle is *not* capacity-matched, and the size of that confound is smaller than the raw counts suggest: channel mixing takes TimeMixer from 179k to 336k parameters at $C{=}21$, but a backward pass shows 133k of the larger model receives no gradient, so the trained capacity ratio is $1.1\times$ rather than $1.9\times$ (at $C{=}321$, $2.0\times$ rather than $2.7\times$). Part of the Weather gain may still be capacity; the sign flip is the more robust observation, since capacity moves the same way on both. Across the three registered datasets the fraction of channels that gain orders as Weather 76%, Traffic 31%, Electricity 11%: monotone in channel count, the axis this literature reaches for, but equally monotone in channel *heterogeneity*, for which a mechanism is available — near-redundant series share structure each channel's own history already carries. Three datasets cannot separate those, and we do not claim a moderator.

5 Was there anything to gain? A closed-form ceiling

Every null in §4 is ambiguous in the same way: when a channel-dependent model shows no gain we cannot tell whether the coupling carried nothing exploitable or whether the model failed to exploit it. "No gain found" is not "no gain available", and that ambiguity is why negative results about cross-channel modeling have been hard to interpret, our own included.

On our generators the question is decidable, and needs no training. All three families are linear with additive Gaussian noise, so conditional on the seed the process is jointly Gaussian and the minimum-MSE predictor from a finite window follows in closed form from the autocovariance sequence via a block-Toeplitz solve. The gap between an all-channel and an own-channel-restricted optimal predictor is the cross-channel value the best model in the correct class could extract. Computed exactly it validates itself, returning -0.00% at $\rho{=}0$ where the channels are independent by construction. What we release as `available_value` is the *estimator* of that quantity, not the exact solve: it costs no GPU and answers the question before a model is chosen, and on these strongly autocorrelated processes its own null reading is -2.9% rather than zero (Appendix D). An earlier version *estimated* this quantity and was wrong to; Appendix D records the null-arm bias correction that was itself the error, inflating diffusion's available value from $+4.0\%$ to $+7.2\%$.

It does not survive contact with real data, and we ship that finding with it. The same computation on Weather returns -368%, which is not a measurement: the method silently assumes stationarity, and a channel nearly flat while the model was fit ($\sigma{=}0.06$) has $\sigma{=}0.86$ at test, so the ridge penalty cannot be selected at all. We make no real-data ceiling claim, restrict the released tool to stationary data, and ship a guard that refuses rather than returning -368% quietly; it checks for a change in per-channel scale between fit and test, not stationarity, so passing it is not a certificate. Three principled repairs all fail, one after looking right at two splits of three (Appendix D).

What the ceiling says, and how much of it we had wrong. For the DepShift topologies the "realized" arm is a matched nested pair — one shared temporal map with and without a channel-mixing layer — so we say "the probe" for those rows. Directed-lag coupling offers $+31\%$ and the probe realizes $+24\%$, about three quarters of it, a share and not a bound (Appendix D), so the oracle does not return zero everywhere; every other configuration also has positive available value — diffusion at $\rho{=}0.8$ offers $+4.0\%$ against -0.6% realized, latent factor $+5.2\%$ against $+3.1\%$. Scored at the final step with no model selection, as our original protocol did, those runs read -2.3% and -3.9% and an earlier version claimed the models did worse than ignoring the other channels; re-scoring at the best held-out step flips six of fifteen configurations. We had called cross-channel information *redundant* under synchronizing coupling. It was not redundant — our models did not use it.

Along the recoverability knob of §4 realized value falls while available value *rises*, so a collapse between two trained models can be theirs rather than the data's. Read without a ceiling it looks exactly like a property of the data, and we read it that way for months (Appendix A).

Applied to the shift, the ceiling validates the protocol and prices equivariance. Fitting the oracles before the break and scoring them after gives an *optimal-predictor* excess fragility of $+38\%$p under permutation and $+24\%$p under a lag change: excess fragility is not an artifact of training. iTransformer's measured excess (2.8%p, 6.0%p) sits 93% below the optimal-predictor cost of a relabeling but only 75% below that of a rewiring. Levels conflate model class and capacity, so

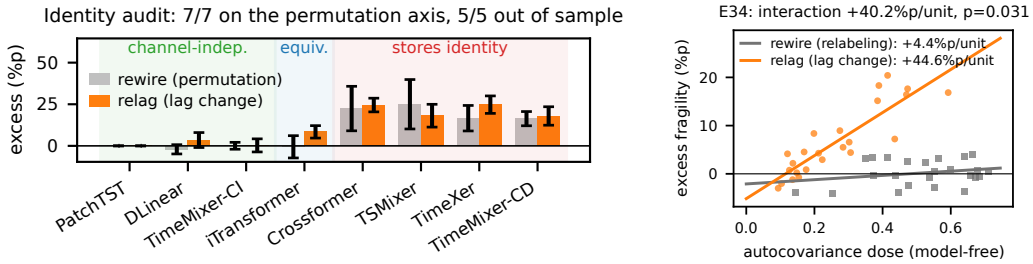


Figure 1: **Left:** the registered audit predicts permutation fragility; only the equivariant arm flips under a lag change. **Right:** that arm on a dose axis in the replication (E34, Appendix Y); lines are the regression's implied slopes.

we rest on the *ordering*, which reverses: at the ceiling a permutation is the more damaging shift ($+14.1$%p paired, Wilcoxon p=0.0001, on 18 of 20 seeds), for iTransformer the less ($+3.2$%p, p=0.0002). Both halves are paired at 20 seeds; at the five we first used, the ceiling half did not reach significance.

The reversal does not show the inductive bias is responsible: a model that never learned the lag structure would reverse it too (Appendix D gives the counterexample, and why iTransformer's lower clean error makes that less likely without excluding it). It does show why a shuffling test flatters such a model: it is graded on the axis the bias was built for.

6 An Architectural Boundary: Who Pays, and For Which Shift

The principle says a cost exists when coupling is non-recoverable; a second, architectural question is *which* models pay it. We *pre-register* predictions from a static audit of each architecture's parameters: a model is predicted fragile to channel permutation iff it stores channel identity in a parameter (a per-channel embedding, positional token, or channel-mixing weight), which breaks permutation equivariance. Committed before running the sweep, the audit gets 7/7 point estimates on the predicted side of the registered thresholds (ROBUST $|\text{excess}| \leq 3$%p, FRAGILE $\geq +10$%p; Fig. 1): channel-independent and variate-equivariant models (DLinear, iTransformer, TimeMixer with channel-independence) show ≈ 0 excess, while identity-storing models (Crossformer, TSMixer, TimeXer, TimeMixer with channel mixing) show large positive excess. TimeMixer, whose channel-independence flag toggles identity storage within one architecture, provides a clean within-model control. So does channel access: iTransformer fed one channel at a time fits 21.9% worse and carries $+1.1$, not $+7.1$%p, lag-change excess (p=0.002; Appendix Y.3).

Four limits on that count. PatchTST's 0 is true by construction and excluded; two more arms were seen in a pilot, leaving 5/5 genuinely out of sample; and at five seeds only 2/7 arms have a 95% CI lying entirely inside the predicted region. What depends on neither is the *ordering*: every arm called fragile exceeds every arm called robust, which on the genuinely out-of-sample arms has exact probability 0.10 — suggestive, not significant, and the number we ask the reader to weigh. A fourth is that the rule does not generalise: on six further architectures, pre-registered from their source, it puts only 3 of 6 on the predicted side, and its one substantive failure stores channel identity without mixing channels (Appendix U).

The causal version of the audit does not go through. Its pre-registered intervention — inject a per-channel identity embedding into iTransformer, replace Crossformer's with a shared one — moves iTransformer as predicted with an interval spanning zero and Crossformer the *wrong* way (Appendix S); identity parameters *predict* permutation fragility; we have not shown they *cause* it.

Is the audit reading architecture or pipeline? Taking evaluation statistics from each window rather than freezing them at training takes iTransformer's own degradation under relabeling flat in k at $+2.06$% by k=8 (E30, Appendix T, which also gives the excess over the control): an exactly equivariant model's error is unchanged by relabeling (Lemma 2), so that residual was the pipeline's

(as in vision; Schneider et al., 2020; Nado et al., 2020). A per-channel parameter is not reached by re-standardising the input, and under the same switch the identity-storing arms keep their fragility whole: Crossformer $+40.4\%$, TSMixer $+48.6\%$, TimeXer $+38.9\%$ at $k{=}8$ against a registered bar of $+4.64\%$, rising with k in every seed (E31; two of the three also clear the registered slope test, so it is reported per architecture). The switch removes the equivariant model's permutation excess and removes none of the identity-storing models'.

Passing a shuffling test is not structural robustness. The audit above concerns *relabeling*. Rewiring is a different disturbance and equivariance does not cover it. We test this with the comparison the claim rests on — the *same* model, the *same* data realization, both shift families, paired by seed at 20 seeds. iTransformer's excess is 2.8% under relabeling and 6.0% under a lag change of the same coupling, a paired difference of $+3.2$%p (95% CI $[+2.0, +4.5]$, Wilcoxon $p{=}0.0002$).

Training the same channel-independent architecture twice gives a *per-seed* |excess| floor of ± 5.0%p that $+3.2$%p does not clear, 15 of 20 per-seed differences falling inside it. But a seed-averaged mean against a single-observation band is not like-for-like: the two null arms' means differ by $+0.01 \pm 0.83$ and the paired flip $+3.2 \pm 0.65$ lies 3.1 combined standard errors from that. We give both; the supported statement is relative, more fragile to rewiring than to relabeling, not badly damaged.

The same holds for the model this argument is aimed at. Testing an equivariant stand-in rather than CPiRi itself is a fair objection, so we ran its released code through the same contrast: the frozen Sundial encoder it ships, and its spatial module, trained by us for six epochs on DepShift. That module is a transformer encoder over channel tokens with no parameter indexed by channel, so it is permutation-*equivariant* by construction, and we verify that on the trained model to 10^{-6} (Appendix H). It behaves accordingly. Across 5 coupling configurations — varying strength and the size *and direction* of the lag change — CPiRi is more fragile to the lag change than to the relabeling in 5 of 5, mean $+10.9$%p, exact one-sided sign test $p{=}0.031$, and on 15 of 15 seeds. The effect tracks how large the lag change is, reaching $+24.7$%p at $\tau\, 4 \to 12$ — the same dose-response the equivariant stand-in shows. Trained with the channel-shuffling regulariser its own paper prescribes for this problem, no closing of the gap is detected: it stays positive at 5 of 5 configurations, the same count and the same $p{=}0.031$ as without it, and the regulariser moves it by -0.4 ± 0.8%p on a $+10.9$%p gap, narrowing it in 2 of 5. That is a failure to detect an improvement, not an equivalence test; an improvement under about two standard errors would not have been seen. The intervention targets the axis that is not the problem.

Five limits, written out in Appendix H: these are raw degradations, not control-subtracted excess; the module trained to a fixed budget without validation selection; every configuration is directed-lag, so $p{=}0.031$ is conditional on that scope; the last two were added after the first three came back positive, making that p nominal; and two share a permutation arm. The unit is the configuration, 3 seeds averaged inside each, for the reason the next paragraph gives.

Seeds are the wrong unit, so we changed it. Everything above is 20 seeds at *one* coupling configuration. Seeds sample initialization and data noise; they do not sample configurations, so that p speaks to run-to-run variation and not to whether the claim holds anywhere else. Running two hundred of them would drive p lower and leave the claim exactly as general as it was. We therefore re-ran the contrast at 12 configurations spanning coupling strength, channel count, the size *and direction* of the lag change, and a second topology, averaged the seeds away inside each configuration, and tested over configurations. The analysis and its falsifiers were committed before the runs finished.

It holds: 11 of 12 configurations are positive, mean $+4.7$%p, one-sided sign test $p{=}0.0032$, descriptive since the configurations share five seeds and two share a permutation arm (Wilcoxon $p{=}0.0012$). Two configurations share a permutation arm, so 11 distinct relabeling measurements sit behind 12 signs. It is not an artifact of the setting we first measured, nor of marginal movement — on every directed-lag configuration the non-permutation shift moves per-channel marginals less than the permutation does (at most $0.97\times$), and on the factor configurations (effect null) at most $1.08\times$.

Nor of disturbance size. Our permutation had one size and relag three, so the contrast could have been one of magnitude. With a size axis for the relabeling too, on one model-free scale, the in-

teraction is $+38.6$%p (positive under every unit and denominator; replicated on six fresh seeds at the seed unit, $+40.2$%p, p=0.0312, Appendix Y) and at a matched dose of 0.52 the fitted values are $+14.3$%p against -0.2%p: the equivariant model is close to indifferent to how far a relabeling moved and is not indifferent to a coupling change (Appendix W, registered). That flatness is not a definition restated: an exactly equivariant model's own error is unchanged by a relabeling (Lemma 2, checked on the trained model in Appendix P), so what its degradation measures above zero is what the pipeline broke — standardisation fitted before the break alone puts a provably equivariant predictor at $+0.17\%$. Against the exact ceiling it is a price not paid: the optimal predictor's excess is $+52$%p at k=8 where iTransformer's is $+0.4$%p, while under a lag change it pays between a tenth and a quarter of the optimal predictor's excess, more at larger doses (Appendix V).

The split across topologies is the more informative half. All 8 of 8 directed-lag configurations are positive (mean $+7.0$%p), and the effect scales with how large the lag change is: lengthening τ from 4 to 12 gives $+16.1$%p and *shortening* it from 8 to 4 gives $+10.5$%p, so it is the size of the structural change and not its direction that matters. In the factor topology, where the non-permutation shift redraws loadings, the effect is small and not significant (mean $+0.3$%p over 4 configurations, 3 positive). This is the second topology our pre-registration required and never had; it does not meet the registered $+10$%p FRAGILE threshold, and we report that rather than the pooled figure alone. One reading is that the effect needs coupling the model exploits: in the factor topology iTransformer's excess is negative under *both* shifts. But that reading infers coupling use from the same excess it is meant to explain, and we have no independent measure of it here, so we state it as a hypothesis and not as the reason for the null.

Out of sample, the ordering holds on both datasets. We tested the audit out of sample, training on an early window and scoring on ten later ones against the channel-independent control, at each dataset's *full* channel count, every arm trained to its own validation optimum, at 3 seeds on Traffic and 7 on Electricity. Excess fragility separates as predicted. On Traffic (C=862) the equivariant arm reads -7.6%p $[-11.5, -3.7]$ against $+7.6$ to $+28.3$%p for all 3 identity-storing arms — every interval excluding zero, and every identity-storing interval disjoint from the equivariant one by $+9.2$ to $+26.3$%p, though not from each other: the claim is about the classes, not a ranking within one. On Electricity (C=321) the equivariant arm reads -3.2%p $[-3.8, -2.5]$ against $+3.2$%p $[+0.5, +5.8]$. The gaps of 15.2 and 6.3%p clear a floor our first version lacked — the same architecture's spread under disjoint seeds, 1.4 and 1.0%p It is the audit's first real-data support not resting on injected coupling.

At the unit that retracted E24. Those intervals are over seeds scoring the same ten windows, the objection Appendix G uses to withdraw a result of ours; at the window unit Traffic's separation survives, "outperforms the control outright" is withdrawn, and Electricity's arm no longer excludes zero (Appendix L).

Electricity carries two caveats Traffic does not, both against us. Its seed count was extended from three, where the surviving arm spanned zero ($[-4.9, +14.3]$), to seven ($[+0.5, +5.8]$): data-dependent stopping. And 2 of the 3 identity-storing arms do not train under the shared recipe and are excluded; one, Crossformer, scores -0.4%p and would be the single arm *contradicting* the audit, so the exclusion removes the most inconvenient point. We keep the rule because both arms train normally on Traffic and failing to train is not fragility (Appendix M).

Attributing it to coupling took three pre-registered attempts, and all three failed. An ordering is not an attribution: these models could be fragile to distribution change in general. A dose-response on the coupling movement between train and test window fired its falsifier, the channel-independent control on Traffic responding as strongly as anything else. A within-architecture contrast meant to cancel that does not survive, because the cancellation argument is *algebraically false* as we stated it, the same error one level down made while writing up the first. A third design addressed both causes and failed at 0.0167; we stopped there, as registered — after overriding, once, a criterion that had already fired (Appendices G, J and C).

So the bounds are these: the effect holds across coupling strength, channel count and lag change on the directed-lag chain and is null in the factor topology, which we read, as a hypothesis we have not measured independently, as needing coupling the model exploits; it persists across C under dense

coupling but attenuates for sparse coupling at high C (Appendix E), and on real data we can show the architectural ordering but not, on 3 registered attempts, that coupling movement drives it.

7 WHERE THE CLAIM STOPS

The pre-registered real-data test is met on one of three datasets, and so fails. We registered the estimator, split, seeds, statistics and pass rule for all three datasets before running, and ran them at that specification with one exception the registration forbade: we changed the estimator. A validity check on the controlled generator showed the registered form scored a single H-step target while the gain is computed over the whole forecast window, leaving it flat across a sweep that varies by construction. We filed a dated amendment before applying the corrected estimator to any real dataset; Weather had already been analysed under the original form, which is why both are reported, and there the as-registered form is *weaker* ($+0.37$ at p=0.10 against $+0.44$ at p=0.046), so the change did not buy the result.

The rule required the predicted (negative) association with a bootstrap CI excluding zero on at least two datasets. Weather gives $+0.44$, the *wrong* sign, with an interval $[+0.39, +0.66]$ that excludes zero under the registered resampling unit — evidence *against* the prediction. Traffic's primary gives $+0.03$ (p=0.398) while its secondary gives -0.18 at p=1e − 07: the two registered estimators disagree in sign, which the rule forbade. Only Electricity satisfies the criterion. Our bootstrap had resampled channel blocks where the registration said time block × seed; at the registered unit no verdict changes and the intervals tighten.

Electricity satisfies the rule for the wrong reason, and the reason is a metric artifact. The criterion is met decisively (-0.74, p=8e−58), but the principle predicts *low*-recoverability channels *gain*, and on Electricity almost nothing gains: only 11% of channels do, median -36.6%. The correlation comes from high-recoverability channels being harmed far more, which is largely the gain's own denominator — easy channels have tiny $\mathrm{MSE}_{\mathrm{CI}}$, so a fixed absolute degradation becomes a huge relative one — the artifact Pearson (1897) identified for indices. In a protocol written by authors trying not to be fooled, it produced the predicted sign at p=2e − 59 and satisfied the rule. Normalizing by target variance shrinks the association to -0.25 and removes its monotonicity across quartiles; Weather becomes a null while Electricity retains -0.25 at p=4×10^{-6}, small, clear, and in the predicted direction. We still decline to read it as support: where almost nothing gains, a correlation orders degrees of harm rather than degrees of benefit.

Where our pre-registrations were not met. Table 1 lists the claims we could not establish. The audit's two-topology requirement went untested until §6's factor arms, which do *not* meet its $+10\%$p threshold; its headline count is five genuinely prospective hits over four architectures, not the eight rows the registration lists; and we declined Electricity's pass on a post-hoc diagnostic. E35, where both shifts fix every single-channel law, was not confirmed: iTransformer gained nothing from the other channels there (Appendix Y.2). One registration was falsified outright rather than unmet, E33's (Appendix U). Appendix C gives each with its multiplicity correction and scope.

Further limitations. Excess fragility divides each arm by its own clean error; over the control's instead, both orderings survive (Appendix K). The synthetic sweeps train to a fixed budget rather than each arm's validation optimum, which inflates magnitudes (Appendix F); the real-data results do select on validation. The contrast is operational, not an identified causal effect, and does not adjudicate model choice: refitting the CD oracle restores its advantage (Appendix D). Our generators are linear-Gaussian, so the ceiling is exact and its nulls do not transfer to nonlinear coupling; the synthetic break is instantaneous where real dependency change is gradual; no single non-permutation family runs across all three topologies; and a fully real *controlled* rewiring is unavailable (Appendix R). Appendix Z has two further negative results.

8 CONCLUSION

A shuffling test certifies only the axis it tests, and the axis it omits is one on which these models are measurably worse. What keeps that from restating equivariance is the generator: its coupling change leaves every single-channel law exactly as it was and no relabeling of these channels can.

ETHICS STATEMENT

This work trains forecasters on public benchmark datasets (ETT, Weather, Electricity, Traffic, Solar, Exchange) and on synthetic series from a generator we release. It involves no human subjects, no personal data, and no human-annotated labels; the Traffic and Electricity series are aggregate sensor and consumption records already published for this purpose, used under the terms they were released with.

The paper's object is an evaluation practice rather than a deployed system, and its result is a negative one: a diagnostic in use certifies less than it is read to certify. We see no dual-use concern in that, and one modest benefit — a practitioner who relies on a channel-shuffling test to decide that a cross-channel model is safe to deploy under changing inter-channel structure is relying on evidence this paper shows does not cover that case.

Two risks are ours to state. First, every result that separates the two disturbances is measured on a linear-Gaussian generator, so a reader who carries the ordering to nonlinear couplings is extrapolating beyond what we tested; §7 says so and Table 1 marks the claims we could not establish. Second, we measure a third party's released method (CPiRi) on an axis their deployment claim names and their diagnostics do not test, and we name it. We report what its released code does under our protocol, quote its own stated scope, and include the configurations where our effect is absent; the finding is about the coverage of a diagnostic, not about the quality of that work.

All experiment code, committed logs, pre-registrations with their dated amendments, and the verification suite are in the supplementary material, and the use of an AI coding assistant is disclosed in full below.

REPRODUCIBILITY STATEMENT

Every number in this paper is regenerated from result JSONs rather than typed: `scripts/reproduce.sh --verify` rebuilds all macro files from the logs and fails on any diff, and `scripts/audit_paper_numbers.py` re-checks every hand-typed number in the paper against the log it comes from, together with relational claims of the form "X and Y do not overlap" that a numeric check alone does not catch, and lists every exemption with its reason. A unit-test suite covers the generator's invariants. The DepShift generator, the closed-form ceiling, and every experiment script are in the accompanying code; external baselines (the Time-Series-Library and CPiRi) are linked read-only and unmodified. The pre-registrations, including their dated amendments and those we failed, are committed as `PREREG_*.md` alongside the runs they govern. Seeds, splits and hyperparameters are recorded in each run's result JSON.

AI USE STATEMENT

We used an LLM-based coding assistant (Claude, Anthropic) extensively and in ways this policy asks us to report. Among the *required* categories: it implemented the DepShift generator and all experiment and analysis scripts; generated the synthetic datasets those scripts define; helped develop the conceptual framework this paper is built on, the relabeling/rewiring split read as two axes of one instrument, and the theoretical model behind the closed-form cross-channel ceiling, which it also derived and implemented; *formulated the two lemmas §3 and §6 rest on and drafted their proofs*, which we then checked numerically on the generator and wired into the verification suite; proposed and refined hypotheses, including the moderator we later retired and the coupling-attribution designs we failed three times; designed experiments and drafted their pre-registrations; and produced first-pass interpretations of results. It did not clean or curate the real datasets beyond the header normalization described in the code, and no translation or qualitative-data analysis was involved. Among the *recommended* categories, it drafted the paper, made the figures and tables, edited code, surveyed related work, and formatted references.

Two things constrain what that assistance was allowed to establish. Every pre-registration, including its pass rule and kill criterion, was reviewed and approved by the authors before the runs it governs, and the outcomes reported here include the ones that failed those rules. And the verification tooling described above exists because assistant-drafted claims did not always survive checking: numeric

claims, relational claims, the page count and several analyses were found wrong on inspection and were corrected or withdrawn, and Table 1 reports the retractions beside the survivors. The authors have checked the claims in this paper against the logs and take full responsibility for its contents.

REFERENCES

Ibram Abdelmalak, Kiran Madhusudhanan, Jungmin Choi, Christian Klötergens, Vijaya Krishna Yalavarthi, Maximilian Stubbemann, and Lars Schmidt-Thieme. Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, volume 16598 of *Lecture Notes in Computer Science*, pp. 585–597, Singapore, 2026. Springer. doi:10.1007/978-981-92-1462-4_46; preprint arXiv:2502.09683v3.

Lorenzo Brigato, Rafael Morand, Knut Joar Strømmen, Maria Panagiotou, Markus Schmidt, and Stavroula Mougiakakou. There are no champions in supervised long-term time series forecasting. *Transactions on Machine Learning Research*, 2026. preprint arXiv:2502.14045v2.

Peter Maurice Catt. Forecastability as an information-theoretic limit on prediction. *arXiv preprint arXiv:2603.27074*, 2026. v1, March 2026.

Ahmad Chamma, Denis A. Engemann, and Bertrand Thirion. Statistically valid variable importance assessment through conditional permutations. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2309.07593v2.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O. Arik, and Tomas Pfister. TSMixer: An all-MLP architecture for time series forecasting. *Transactions on Machine Learning Research*, 2023. September 2023; preprint arXiv:2303.06053v5.

Wanjin Feng, Yuan Yuan, Jingtao Ding, and Yong Li. Beyond model ranking: Predictability-aligned evaluation for time series forecasting. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2509.23074v3.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Fernando Gama, Joan Bruna, and Alejandro Ribeiro. Stability properties of graph neural networks. *IEEE Transactions on Signal Processing*, 68:5680–5695, 2020.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

Lu Han, Han-Jia Ye, and De-Chuan Zhan. The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7129–7142, 2024. doi:10.1109/TKDE.2024.3400008; preprint arXiv:2304.05206.

Hui He, Qi Zhang, Kun Yi, Xiaojun Xue, Shoujin Wang, Liang Hu, and Longbing Cao. Robust multivariate time series forecasting against intraseries and interseries transitional shift. *IEEE Transactions on Neural Networks and Learning Systems*, 36(12):20357–20370, 2025. Preprint arXiv:2407.13194.

Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):82, 2021. arXiv:1905.03151v2.

Jinkwan Jang, Hyungjin Park, Jinmyeong Choi, and Taesup Kim. Towards robust real-world multivariate time series forecasting: A unified framework for dependency, asynchrony, and missingness. In *International Conference on Learning Representations*, 2026. arXiv:2506.08660v4.

Seungwoo Jeong and Heung-Il Suk. Permutation-equivariant 2D state space models: Theory and canonical architecture for multivariate time series. *arXiv preprint arXiv:2603.08753*, 2026. v1, March 2026.

Henry Kenlay, Dorina Thanou, and Xiaowen Dong. On the stability of graph convolutional neural networks under edge rewiring. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8513–8517, 2021. arXiv:2010.13747v2.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. v3, January 2021.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Karl Pearson. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498, 1897.

Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Junkai Lu, Jilin Hu, Chenjuan Guo, Christian S. Jensen, and Bin Yang. A comprehensive survey of deep learning for multivariate time series forecasting: A channel strategy perspective. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26), Survey Track*, pp. 7995–8003, 2026. doi:10.24963/ijcai.2026/887; preprint arXiv:2502.10721v3.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2006.16971v2.

Qitai Tan, Yiyun Chen, Mo Li, Ruiwen Gu, Yilin Su, and Xiao-Ping Zhang. SynTSBench: Rethinking temporal pattern learning in deep learning models for time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 161198–161242, 2025. doi:10.52202/085713-4860; arXiv:2510.20273v1.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations*, 2022. arXiv:2111.14522v3.

Annita Vapsi, Penghang Liu, Saheed Obitayo, Aakriti, Manoj Cherukumalli, Prathamesh Patil, Amit Varshney, Nicolas Marchesotti, Elizabeth Fons, Vamsi K. Potluru, and Manuela Veloso. Dynamic linear coregionalization for realistic synthetic multivariate time series. In *ICLR Workshop on Time Series in the Age of Large Models*, 2026. arXiv:2604.05064v2.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024a. arXiv:2405.14616v1.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 469–498, 2024b. doi:10.52202/079017-0015.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations (ICLR)*, 2022.

Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022. arXiv:2202.02466v5.

Jiyuan Xu, Wenyu Zhang, Xin Jing, Shuai Chen, Shuai Zhang, and Jiahao Nie. CPiRi: Channel permutation-invariant relational interaction for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2601.20318v2.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. issue 9; doi:10.1609/aaai.v37i9.26317.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2023.

Lifan Zhao and Yanyan Shen. Rethinking channel dependence for multivariate time series forecasting: Learning from leading indicators. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2401.17548v6; code https://github.com/SJTU-DMTai/LIFT.

A THE RETIRED MODERATOR IN FULL

This appendix carries the construction, the per-seed table and each falsification of the candidate moderator summarized in §4; Table 2 gives the per-seed detail. It is here rather than deleted because the arc — a clean co-movement, three failed generalisations, and finally a ceiling computation showing the co-movement was never in the data (§5) — is the part of this work we would most want another group to be able to check. We construct a controlled recoverability knob: parents are AR(1) processes with coefficient ρ_p; followers are lagged copies of a parent plus noise, so a follower's own-lag-1 autocorrelation equals ρ_p (verified). Sweeping ρ_p from 0.1 to 0.9 — which moves measured recoverability from ≈ 0 to only 0.16, so the sweep spans "unrecoverable" to "weakly recoverable", not to recoverable — over 8 seeds, Fig. 2 shows the cross-channel *value* and the dependency-shift *fragility* (excess degradation under a coupling rewiring) fall toward zero *together* ($9.6\%/7.4\%$ to $0.7\%/0.3\%$). One data property moves both. Because our own analysis warns that ratio metrics can be driven by their denominator (§7), we checked this one: across the sweep the CI control's error moves by 7.5% while the absolute gap $\mathrm{MSE}_{\mathrm{CI}} - \mathrm{MSE}_{\mathrm{CD}}$ closes by 92% ($0.108 \to 0.009$), and the trend is equally monotone in the absolute gap. Here the effect is in the numerator.

The post-mortem: the moderator's trend was a property of the models. On the construction that motivated the moderator retired in Appendix A, realized value at final-epoch scoring, the rule this section later retires, falls from $+9.6\%$ to $+0.7\%$ as own-history recoverability rises (at best-validation scoring, $+7.4\% \to +5.9\%$, the corrected figures three paragraphs down) — the co-movement that made the hypothesis attractive. Available value does the opposite, rising $+6.8\%$ to $+9.9\%$. Over the 5 sweep points Spearman is $+1.00$ between the knob and available value (exact permutation p=0.017) and -0.90 between the knob and realized value (p=0.083). *We rest the claim on the available arm*, which is the powered one and is the arm that does the work: the retired moderator predicted cross-channel value *falls* as recoverability rises, and available value rises instead. The

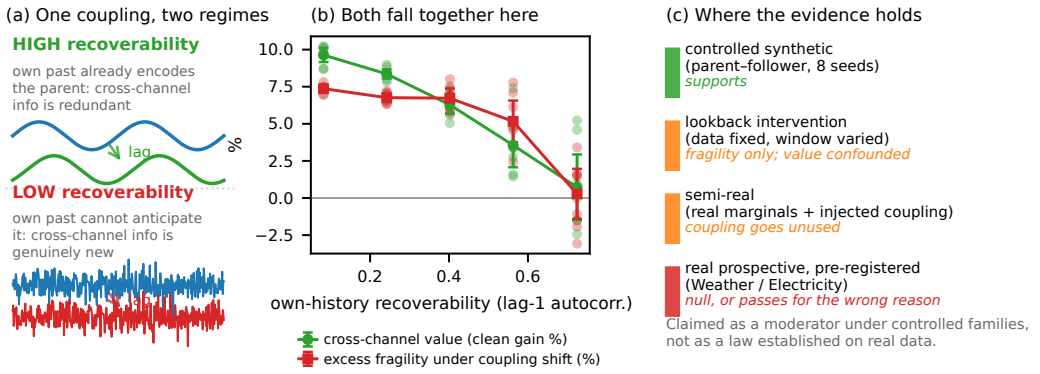


Figure 2: **(a)** The same cross-channel dependency under low vs. high own-history recoverability. **(b)** In the directed-lag construction, cross-channel value and excess fragility fall together as measured recoverability rises (mean $\pm$95% CI, 8 seeds). The abscissa is the follower channels' lag-1 autocorrelation; the pre-registered AR R^2 estimator runs over a much narrower range on this sweep (-0.01 to 0.16) but orders the configurations identically, so the trend does not depend on which readout is plotted. **(c)** Where this does and does not hold: the lookback intervention reproduces only the fragility trend, an independent generator family does not reproduce it, and the pre-registered real-data tests do not support it.

Table 2: The directed-lag construction in full, so the retired hypothesis can be inspected rather than taken on trust: measured recoverability, clean MSE for the CI and CD arms, the resulting value, each arm's degradation under the coupling rewiring, and the excess. Mean $\pm$95% CI over 8 seeds. This is the one construction in which value and fragility descend together; the generalization tests below do not reproduce it.

		clean MSE		shift degradation			
ρ_p	recov. (AR R^2)	CI	CD	value	CI	CD	excess fragility
0.1	-0.008	1.118 ± 0.016	1.010 ± 0.013	$9.6\% \pm 0.5$	$-0.5\% \pm 1.7$	$6.9\% \pm 1.6$	$7.4\% \pm 0.2$
0.3	0.001	1.127 ± 0.018	1.032 ± 0.017	$8.4\% \pm 0.3$	$-0.5\% \pm 2.0$	$6.3\% \pm 2.0$	$6.8\% \pm 0.3$
0.5	0.019	1.156 ± 0.021	1.084 ± 0.019	$6.3\% \pm 0.6$	$-1.1\% \pm 2.5$	$5.6\% \pm 2.6$	$6.7\% \pm 0.6$
0.7	0.049	1.202 ± 0.030	1.159 ± 0.021	$3.6\% \pm 1.5$	$-1.7\% \pm 3.6$	$3.5\% \pm 3.1$	$5.2\% \pm 1.4$
0.9	0.163	1.163 ± 0.036	1.154 ± 0.039	$0.7\% \pm 2.2$	$-2.9\% \pm 4.7$	$-2.7\% \pm 4.7$	$0.3\% \pm 1.7$

realized arm points the other way but at n=5 sweep points cannot reach significance — its lowest attainable Wilcoxon p is 0.0625 — and it is non-monotone at the last point. Presenting the opposition as though both arms were established would repeat the error this section is about.

One of our own protocol choices was inflating this, and we corrected it. The sweep as first run scored each model at its *final* epoch with no model selection. Re-scoring the same runs at their best validation epoch — same data, seeds and budget, selection on a held-out slice of train and never on the split we report — leaves the direction intact but shrinks the magnitude by an order of magnitude: realized value declines $+7.4\% \to +5.9\%$ rather than $+9.3\% \to +0.3\%$, and at the top of the sweep the networks capture 64% of what was available, not 3%. The supported statement is the weaker one: the networks realize *less* of the available value as own-history recoverability rises, not almost none of it. We report the corrected figures and flag this because it is our own thesis one level down — we came close to attributing to the models what was partly a property of when we stopped training them.

Putting the four arms on one scale (Table 5, appendix) explains it, and the explanation is worse for the field than the discrepancy was. At ρ_p=0.9 the achievable frontier *improves* (CI ceiling $1.011 \to 0.827$) and available value *rises* to $+9.2\%$, yet both networks get *worse* in absolute terms, ending 41% and 40% above the frontier. The realized "value" at final-epoch scoring, which Table 5 reports, collapses to $+0.7\%$ not because there was less to gain but because two models far off the

frontier converged to each other. At ρ_p=0.1, likewise, iTransformer's headline $+9.6\%$ "gain" is it beating a baseline that is 11% worse than the best linear CI predictor, while itself landing *on* that CI ceiling — capturing none of the $+6.7\%$ available to it.

The lesson is not about these two architectures. A relative comparison between two trained models can collapse purely because both degrade toward each other, with no change — or an increase — in what was achievable. Read without a ceiling that looks exactly like a property of the data, and we read it that way for months.

This did not generalize, and we report that. We treated the co-movement as a hypothesis and tried to break it, pre-registering the tests where a real-data claim was at stake. It did not survive. (i) A *lookback intervention*, which raises recoverability while holding the data fixed, reproduces the fragility trend but not the value trend: value plunges to -60% at long windows, because lengthening the window also changes how well each variant fits, so the intervention cannot isolate value. (ii) An *independent generator family*—a latent factor whose recoverability is moved by idiosyncratic noise, with the factor's autocorrelation held fixed—does not reproduce it: Spearman is -1.00 for value but -0.40 for fragility, and, more tellingly, the largest value anywhere in that family is $+0.5\%$, so there was never a gain for recoverability to moderate. (iii) The real-data tests behave as described in §7. We therefore do *not* claim recoverability as a general moderator. What survives is narrower and still useful: in this construction, one data property moves both quantities, which is a concrete instance of the broader pattern that cross-channel gains and their shift exposure are not independent knobs.

B EVERY CLAIM, WITH ITS EXPERIMENT, LOG, UNIT AND STATUS

Table 3 has one row per claim the body makes, so a reader can find the run behind a sentence without reconstructing it. Registrations are the `PREREG_*.md` files in the artifact; logs are under `results/logs/`; the unit is the independent block the reported test treats as exchangeable.

C PRE-REGISTRATION DEVIATIONS, IN FULL

The two registered estimators are not interchangeable. They agree at $r = +0.96$ on Weather but only $r = +0.32$ on Electricity and $r = +0.23$ on Traffic. We had called them measures of the same construct on the strength of the Weather number alone. On hundreds of near-redundant series, "how well does a channel's own past predict it" by a model-free autoregression and by a trained CI model's normalized error come apart, and on Traffic they disagree in sign. A pass rule whose two estimators can diverge that far was weaker than we believed when we wrote it.

Summarised in §7; given here without abbreviation. We invoke pre-registration as a discipline, so we list its failures rather than only its successes. (i) The architecture audit registered the FRAGILE criterion as "excess $\geq +10$%p on ≥ 2 of the complementary topologies"; we ran one (directed-lag) for the audit sweep. We do have factor-topology runs for a subset of arms, but a naming defect in our generator means they do not test the registered prediction: in the factor topology the mode we called `rewire` *redrew* the loadings rather than permuting them, so those runs are a non-permutation structural shift filed under the permutation name. The prediction being audited is about *permutation* fragility, so the registered two-topology requirement is simply unmet—not met and disconfirmed, but untested in the second topology. For completeness the numbers under that (mislabeled, non-permutation) shift are Crossformer excess $+1.9$, $+15.7$, $+6.9$, $+12.3$ across four noise levels. We have since separated the two modes in the generator (`rewire` = permute loadings, `reload` = redraw them) and added a contract test asserting that `rewire` preserves the multiset of coupling parameters in every topology; the corrected factor sweep is not in this version. (ii) The headline count is not eight independent prospective hits (the registration lists eight rows). The CI control is zero by construction, three of the eight rows were already observed when the registration was written (the document says so), and "eight architectures" is seven, since TimeMixer appears twice via its channel-independence flag. The genuinely prospective, non-tautological count is five, over four architectures. (ii-b) *A kill criterion fired and we overrode it.* After the first attribution attempt (E23) failed, our own plan's stop rule said to stop pursuing this venue. We did not stop; we ran a second attempt and then a third. `PREREG_E24.md` records the override in its opening section rather than absorbing it quietly, and commits the paper to reporting it, which this paragraph does.

Table 3: Every claim the body makes, with the experiment behind it, its log, the independent unit its test treats as exchangeable, and its status.

claim	registration	log	unit	status
rewiring hurts more than relabeling, 12 configs	E20	`phase_e20`	configuration (descriptive; shared seeds)	supported
same, on CPiRi's released code	E26	`phase_e26`	configuration, sequential (n=5, nominal)	supported
dose interaction, relag $\gg$ rewire	E27	`phase_e27`	seed (G=3; floor 0.25)	coefficient positive; not established at this unit
same, on six fresh seeds	E34	`phase_e34`	seed (G=6)	supported, confirmatory
both axes, every single-channel law fixed, dense factors	E35	`phase_e35`	seed (G=10)	not confirmed; iTransformer gained nothing from the other channels
lag-change excess comes from channel access, same model	E36	`phase_e36`	seed (G=10)	supported, confirmatory
exact reliance reads use without retraining	E37	`phase_e37`	seed (G=10)	supported (instrument check)
identity audit, 7 arms	E2	`phase_e2`	arm (5 prospective)	suggestive
audit on 6 unseen arms	E33	`phase_e33`	arm	falsified
audit survives adaptive normalisation	E31	`phase_e31`	arm, 4-point slope (3 seeds)	supported, per arm
causal version of the audit	E1	`phase_e1`	seed	not shown
ordering out of sample, full C	E25	`phase_e25`	seed; window (Appendix L)	supported (Traffic); sign only (Electricity)
real-data attribution to coupling	E23, E24, E29	`phase_e23`, `e24`, `e29`	window	failed three times
lag change holds single-channel laws	Lemma 1	`logs/theory`	exact	supported (single parent)
marginal-preserving relabeling	E32	`phase_e32`	cell	void by G3
ceiling: value scarce, not absent	E17/E11	`phase_e17/ e11`	exact / estimate	supported
ceiling transfers to real data	E12, E18	`phase_e12/ e18`	—	not shown
recoverability moderates value	RECOVER-ABILITY	`phase_e6/e7`	sweep point	retired

The consequence is the one that registration names: the attribution question was tested three times, not once, so the two later designs carry a multiplicity the first did not, and "we stopped there, as registered" describes only the third stop. The registered stop we honoured is worth less as evidence of discipline than the one we did not. (iii) The real-data protocol registered *three* datasets at five seeds and all three were run at that specification, including Traffic (C=862), whose verdict we could have predicted and ran anyway. An earlier version of this paper reported Weather from a three-seed pilot and quoted its *secondary* estimator while describing it as the primary; that is corrected above. (iv) We report each association without a multiplicity correction. Applying Holm at α=0.05 across the six tests (three datasets $\times$ two registered estimators) leaves both Electricity results and the Traffic secondary significant, and removes Weather's nominal p=0.046, which we already decline to read as support because its sign is wrong. The correction changes no verdict among those six; we state it so a reader need not work out whether it would. It covers that family only. The tests the headline rests on — E20's twelve-configuration sign test, the dose plane's four falsifiers, CPiRi's five configurations (two added sequentially), E31's three arms on two criteria, and H3/H5/H6 registered together — are each reported at their own registered threshold with no correction across them. The architecture audit's seven arms, the ceiling sweep's fifteen configurations, the recoverability sweep's five points, and the registered real-data attribution attempts of §6 — both of which failed, so no correction could

rescue them — are four further families of tests reported here uncorrected, and a reader weighing any single p from them should discount accordingly. We do not correct across them because they test different predictions on different data and pooling them into one family would be a choice as arbitrary as ignoring multiplicity altogether — but the reason we can say the correction changes nothing is that our conclusions rest on orderings and on paired contrasts, not on any individual threshold being cleared. (v) When Electricity did meet the registered criterion, we declined the pass on a post-hoc diagnostic (above). We think that call is right, but taken together these mean the pre-registration constrained us less than the word implies, and a reader should weight it accordingly.

(v) The same-class floor in §6 was first printed as half the twin's 95% interval, which is a standard error and shrinks with seeds: extending Electricity from three to seven seeds took it from 1.9 to 0.9%p without our saying so, a change in the authors' favour that a reviewer found. The floor is now the twin's seed SD, or its mean offset where larger (1.0%p on Electricity, 1.4 on Traffic); both definitions and both values are in the log, and no verdict changes.

D ORACLE CALIBRATION, NULL ARMS AND COMMENSURABILITY

Unpenalized least squares at full lag depth overfits pure noise by roughly 14% at our sample sizes, which would fabricate a *negative* available value out of nothing. We therefore select both the lag depth and a ridge penalty on a held-out slice of the training segment, separately for each model class, so neither side is handicapped. Three cases with known answers check the result: independent white noise and independent AR(1) processes both read ≈ 0 ($+0.09\%$, -0.18%), and a channel that is an exact lag-4 copy of another reads $+50.1\%$ against a true $+50\%$. Even so, on the strongly autocorrelated DepShift processes the estimator reads -2.9% where the true value is exactly zero by construction (ρ=0, no coupling). We do not hide this behind the calibration cases: every family has a coupling-free null arm, and Table 4 reports the exact value beside the estimate so the bias is visible per row. An earlier version subtracted a bias measured on that null arm; the paragraph after the table records why that correction was wrong and was retired, and for DepShift the ρ=0 null does not even preserve the per-channel scale, so it is not a marginals-matched control. Levels below a few percent should be read as indistinguishable from zero. Trends across a sweep are readable only where the bias is constant along it, which the per-configuration null readings below show it is not everywhere.

We arrived at it the hard way. The estimator overfits pure noise by $\approx 14\%$ at full lag depth, so we selected lag depth and ridge penalty on held-out data and then subtracted a bias measured on a coupling-free null arm. The estimator turned out to be nearly unbiased where it was actually used (within 0.4%p on every DepShift row) and the *correction* was the problem: the DepShift null is the ρ=0 configuration, which is the most ill-conditioned point in the family — an undiluted r=0.97 AR(1), per-channel σ=2.35 against 1.43 at ρ=0.8 — so a bias measured there does not transfer. Subtracting it inflated diffusion's available value from a true $+4.0\%$ to a reported $+7.2\%$, an 81% overstatement, and made a *bias-corrected point estimate* exceed the ceiling it claimed to be. The general lesson is not specific to us: a surrogate that destroys cross-channel structure necessarily changes the conditioning of the channel-dependent design, which is the sole source of the bias it is meant to measure.

Three attempts to make the ceiling work on real data, and why none ships. The restriction to stationary data is the most limiting thing about this diagnostic, so we tried to remove it.

(i) Standardise each segment by its own mean and standard deviation. The guard blames a scale mismatch, so remove the scale. On a controlled pure scale shift this is exactly right: eight independent AR(1) channels whose true value is 0, with one channel scaled down on the train segment only, give -0.46% at every shift size we tried while the unmodified estimator degrades from -0.5% to -40.7% as the shift grows. On Weather it is far *worse* than doing nothing. Both results are real, and together they say the guard's diagnosis is incomplete: Weather's failure is not a scale shift.

(ii) Standardise with a robust scale. Weather's real pathology is heavy tails — one channel has kurtosis 5.3×10^4, and its standard deviation is 1.18 on one segment against 0.027 on another purely according to whether a spike fell inside the window. For such a channel the standard deviation is not a scale at all. Median/MAD is the natural repair and is the worst of the three: the MAD of a spike channel is tiny, so dividing by it amplifies the spikes instead of taming them.

Table 4: Available versus realized cross-channel value. The oracle is a ceiling here: the generators are linear-Gaussian, so least squares is the correct model class. The two Available columns are the exact block-Toeplitz computation and the released estimator with its held-out interval; the retired null-corrected column is not shown. The two right-hand columns are commensurable only for the DepShift rows, where the realized number comes from a matched nested pair; elsewhere it contrasts two architectures and carries an architecture offset (Appendix D). ‡Two distinct families are named "factor" and they are not the same construction: "Latent factor (DepShift)" is a DepShift topology scored by the matched nested pair, while "Factor family (E9)" is the independent generator of Appendix A's falsification tests scored by a two-architecture contrast. That is why their realized columns disagree in sign — the E9 rows carry an architecture offset that at low idiosyncratic noise reaches -20%, and it is the offset rather than a failure to use the coupling. Only the DepShift rows should be read against the ceiling.

Configuration	Available (%)		Realised (%)
	exact	estimated	by networks
Diffusion, ρ=0.2	$+3.7$	$+4.1 \pm 2.7$	$+0.5$
Diffusion, ρ=0.8	$+4.0$	$+4.3 \pm 2.0$	-0.6
Directed-lag, ρ=0.2	$+19.7$	$+19.8 \pm 0.4$	$+16.2$
Directed-lag, ρ=0.8	$+31.1$	$+31.4 \pm 1.5$	$+23.8$
Latent factor (DepShift), ρ=0.2	$+1.0$	$+0.9 \pm 0.6$	-0.4
Latent factor (DepShift), ρ=0.8	$+5.2$	$+5.6 \pm 3.9$	$+3.1$
Parent–follower, ρ_p=0.1	$+6.7$	$+6.7 \pm 0.1$	$+9.6$
Parent–follower, ρ_p=0.3	$+6.9$	$+6.9 \pm 0.1$	$+8.4$
Parent–follower, ρ_p=0.5	$+7.2$	$+7.2 \pm 0.1$	$+6.3$
Parent–follower, ρ_p=0.7	$+7.8$	$+7.7 \pm 0.2$	$+3.6$
Parent–follower, ρ_p=0.9	$+9.8$	$+8.9 \pm 0.9$	$+0.7$
Factor family (E9)‡, σ_e=2.0	$+2.2$	$+2.1 \pm 0.6$	$+0.5$
Factor family (E9)‡, σ_e=1.0	$+3.8$	$+3.8 \pm 0.6$	-8.4
Factor family (E9)‡, σ_e=0.5	$+4.0$	$+3.7 \pm 0.5$	-12.9
Factor family (E9)‡, σ_e=0.25	$+2.6$	$+2.1 \pm 1.0$	-20.0

(iii) Cap the CD lag depth. Weather's channel-dependent design is $L \times C = 2016$ columns and the selection is plainly overfitting it. Capping the depth at 24 lags gives $+2.2\%$ and $+2.5\%$ at the first two splits we tried — plausible, stable, and the result we would have reported. At the third split it gives -3.3%. A quantity that changes sign with where the split falls is not a ceiling.

So the guard stays, and the honest statement is the one we make in §5: this estimator has a validated domain and real non-stationary data is outside it. We report the attempts because "we did not try" and "three principled repairs each fail for a different reason" are different limitations, and because (iii) is a reminder that two splits agreeing is not stability.

Our metric was never a verdict on model choice. Refitting the CD oracle on the post-break regime — an upper bound on the test-time re-estimation LIFT (Zhao & Shen, 2024) performs — recovers only 25% of the error a permutation added, yet restores the CD advantage ($+29\%$ clean, $+4\%$ shifted, $+29\%$ refitted), and the CD oracle still wins on 14 of 20 seeds after a permutation even unadapted. So: *large excess fragility does not mean the channel-dependent model is the wrong choice*. The metric prices a dependency change; it does not adjudicate model selection.

Null arms. Each family's coupling-free arm cuts the cross-channel path; only E6's and E9's also preserve the per-channel marginals, and DepShift's does not: E6's followers are driven by an independent copy of the parent process, E9's channels each get their own latent factor, and the DepShift topologies use their own ρ=0 configuration. The measured bias tracks how autocorrelated the channels are — -0.1% on the parent–follower family at low ρ_p, -1.1% at high ρ_p, -1.0 to -2.0% across the factor family, and -2.9% on DepShift at r_{slow}=0.97.

On the withdrawn model-class argument. We had argued from this that a difference in model class cannot flip which of two shifts hurts more, so the reversal must be the inductive bias. That

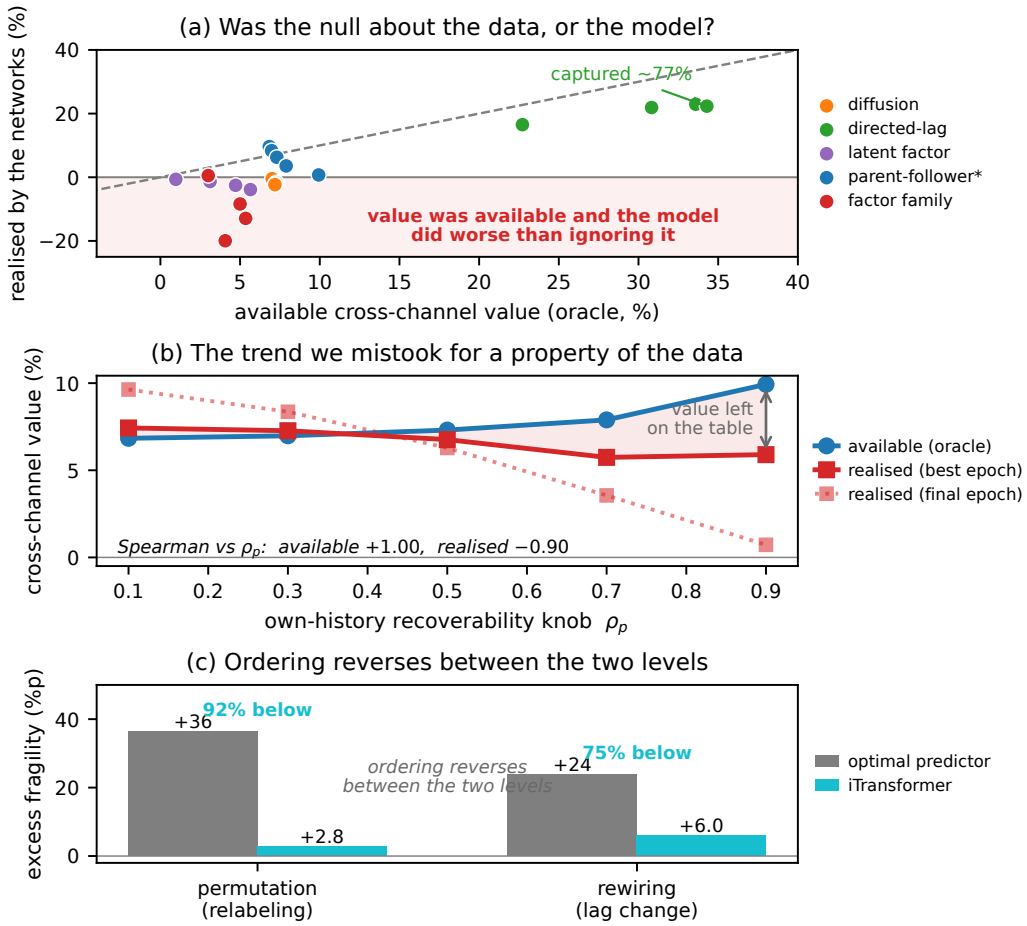


Figure 3: **(a)** What a channel-dependent model realized against what was available; the diagonal is where the two would be equal. Only directed-lag coupling (green) puts a large gain on the table, and there the realized gap is about three quarters of it — a descriptive ratio of two models' errors, not a measured capture rate, since neither network is shown to reach its class optimum. **(b)** The sweep that motivated the moderator we retired: realized value declines while available value *rises* (Spearman $+1.00$ against -0.90). Without the blue curve the red one looks like a property of the data. The dotted line is the same runs scored at their *final* epoch, as our original protocol did — most of the apparent collapse is late-training degradation. **(c)** Excess fragility between optimal predictors against what iTransformer pays; the *ordering* reverses between the two levels. Panels use each experiment's own selection protocol and the parent–follower points carry an architecture offset, so they are comparable by trend rather than by level (Appendix D).

argument is wrong and we withdraw it: a model that never learned the lag structure cannot be hurt by changing it, so its relag excess would sit near zero while its permutation excess did not — an ordering flip produced by hypothesis class alone, with no equivariance involved. What we can say is narrower. That particular alternative is made less likely here, not excluded: iTransformer's clean error is 22% below the channel-independent control's in this configuration, which is consistent with its using the coupling it is later asked to survive a change of; a cross-architecture gap carries an architecture offset, so it is evidence and not proof on its own; the same-model ablation of Appendix Y.3 settles it for iTransformer on the chain. The general form of the objection is not excluded, and the measurement that would exclude it is a within-architecture one, such as TimeMixer's channel-independence flag, which toggles channel access within one architecture; that changes the hypothesis class too, so it narrows the confound rather than removing it. We name that as the next experiment rather than claiming its result.

Table 5: All four arms on one scale (parent–follower, clean MSE; networks at final-epoch scoring, the rule Appendix A retires, so this table shows the uncorrected magnitudes). Each network's excess over its own class optimum is $+11\%/+7\%$ at ρ_p=0.1 and $+41\%/+54\%$ at ρ_p=0.9 (CI/CD), which is what the ratio of realized to available gain does not show. As ρ_p rises the achievable frontier improves and available cross-channel value rises, while both networks move away from the frontier. The relative "value" between them collapses anyway.

	Oracle (achievable)		Networks		Cross-channel value	
ρ_p	CI	CD	PatchTST	iTransformer	available	realized
0.1	1.011	0.944	1.118	1.010	+6.7%	+9.6%
0.9	0.827	0.751	1.163	1.154	+9.2%	+0.7%

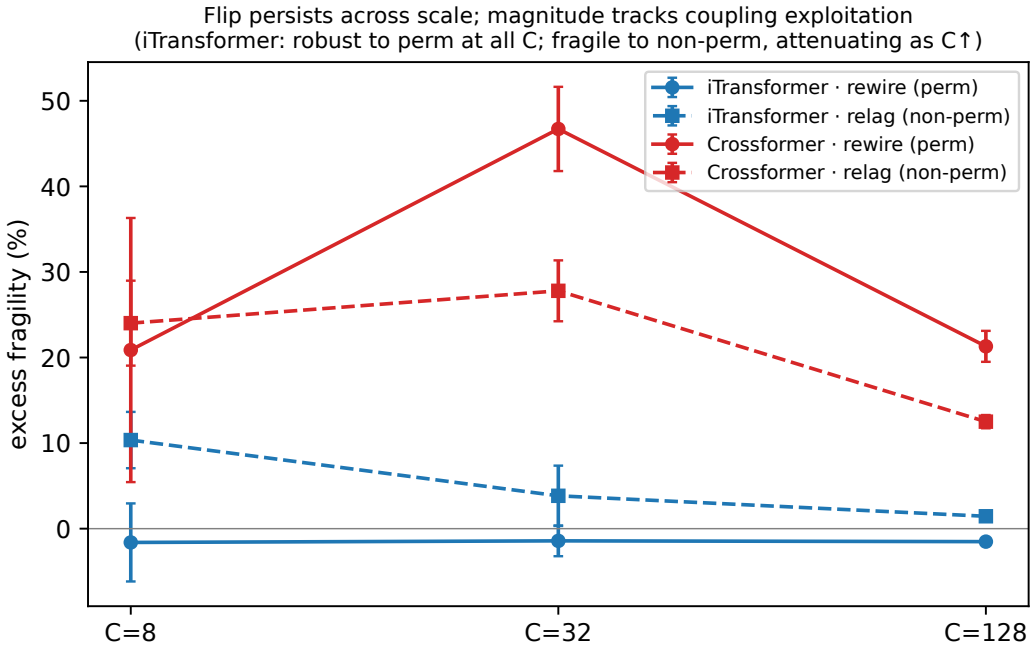


Figure 4: The rewiring effect across channel count; magnitude tracks coupling exploitability.

Commensurability. The realized column is commensurable with the ceiling only where the realized number comes from a matched nested pair — one shared temporal map with and without a channel-mixing layer, as in the DepShift rows. For the parent–follower and factor families it contrasts two different architectures, so it carries an architecture offset as well as a channel-access effect and can exceed the ceiling, which is what happens at ρ_p=0.1. Table 5 puts all four arms in absolute MSE for that reason. We had claimed trends survive this offset because whatever separates the two definitions is a property of the architectures rather than of ρ_p. That is false, and our own numbers say so: the measured offset runs $+0.8, +0.4, -0.5, -2.0, -3.3$ across the sweep — monotone in ρ_p, and $2.7\times$ the entire realized decline. The trend comparison is therefore not rescued by an appeal to a constant offset, which is why we rest the conclusion on the available arm alone.

E THE REWIRING EFFECT ACROSS CHANNEL COUNT

Referenced from §6 (Fig. 4). The effect's magnitude tracks how much coupling the model actually exploits: it persists across C under dense coupling and attenuates for sparse coupling at high C. The dense arm gives each follower several parents, and Appendix P shows the lag change then moves per-channel variances by up to 15.3%; so on that arm the lag change is a coupling change *and* a marginal one, and its excess is read against the control with that caveat.

Table 6: The same two mechanisms at *equal* budget on Weather (three seeds, patience = epochs = 30, selection at best validation epoch). Compare Table 7, where the backbone was allowed roughly twice the search.

Mechanism	ΔMSE, H=96	ΔMSE, H=336
backbone (dep=none)	0.1717	0.2652
median epoch to best	*2*	*2*
CSCA	+5.1%	+2.7%
median epoch to best	*1*	*1*
SD-Graph	+0.4%	+0.5%
median epoch to best	*2*	*2*

Table 7: Cross-channel mechanisms vs. the backbone (ΔMSE; negative = improvement) on high-channel data. Median epochs to best validation loss are shown because the arms are *not* equally trained — the backbone runs roughly twice as long before early stopping fires. † marks a cell whose runs all peaked at epoch 1, measuring optimization failure rather than the mechanism. MoW is listed separately: its gate is per-channel over wavelet bases and carries no cross-channel information.

Mechanism	Weather	Electricity	Traffic
backbone (dep=none)	0.213	0.153	0.402
median epochs to best	*14*	*12*	*14*
Cross-channel mechanisms			
CSCA (Δ vs backbone)	+10.4%	+9.4%	+8.2%
median epochs to best	*5*	*5*	*8*
SD-Graph (Δ vs backbone)	+3.6%	+0.2%	+2.6%
median epochs to best	*6*	*12*	*6*
Not a cross-channel mechanism (per-channel gate over wavelet bases)			
MoW (Δ vs backbone)	-2.8%	+30.2%†	+25.5%
median epochs to best	*8*	*4*	*5*

F BOLT-ON CROSS-CHANNEL MECHANISMS, UNEQUALLY TRAINED

Referenced from §4. Two cross-channel mechanisms added to a strong backbone — scale-level variable attention (CSCA) and a scale-stratified correlation graph (SD-Graph) — read as net-negative on all three high-channel benchmarks in Table 7. That table is *not* evidence, because its arms are not equally trained: they stop at 5–8 epochs against the backbone's 12–14 under patience 3, so an unknown part of each Δ is optimization rather than mechanism.

We have since run the experiment that settles it. Re-running on Weather with patience set equal to the epoch budget, so every arm searches all 30 epochs while selection stays at the best validation epoch, roughly halves CSCA's penalty and all but removes SD-Graph's (Table 6): CSCA goes from $+10.4\%$ to $+5.1\%$ and SD-Graph from $+3.6\%$ to $+0.4\%$ at H=96. So the unequal budget overstated the negative result by roughly a factor of two for one mechanism and nine for the other, and at equal budget SD-Graph is *neutral* rather than harmful. What does survive is specific to CSCA: it still peaks at epoch 1 with thirty available, so that arm fails to train for a reason early stopping was not causing. Separating those two things is the point — the original table bundled a real optimization failure with a budget artifact and reported the sum as evidence about cross-channel modelling.

A third arm we previously counted here (a mixture of lifting-wavelet bases) is *not* cross-channel at all — its gate is per channel over bases — and is excluded.

G TWO REGISTERED ATTEMPTS AT REAL-DATA COUPLING ATTRIBUTION, AND WHY BOTH FAIL

Referenced from §6. Both used the same windowing — train on the first 35% of Electricity and Traffic, a clean holdout to 40%, ten dose windows over the remainder — and the same doses, the

coupling and marginal movement between the training window and each later window measured on the raw series.

First attempt: absolute degradation. We regressed each arm's $\delta(w) = (\mathrm{MSE}(w) - \mathrm{MSE}_{\text{clean}})/\mathrm{MSE}_{\text{clean}}$ on the two doses, registering $b_{\text{coup}} > 0$ for the identity-storing and equivariant arms and, as the falsifier, that it is *not* positive for the channel-independent control, which cannot use coupling at all. On Electricity the coefficient is null — raw dose/degradation correlations of -0.02 to -0.08, $R^2 \leq 0.09$. On Traffic the falsifier fires: the control reads $+1136$ ($p < 0.0001$) with every other arm positive beside it, and with the two doses uncorrelated there (r=0.03) collinearity cannot be blamed.

Second attempt: a within-architecture contrast. The diagnosis names a repair: make the outcome $\delta_{\text{CD}} - \delta_{\text{CI}}$ for one backbone, so a response common to both arms cancels. We added a capacity control (the channel-independent arm widened past the mixing arm's parameter count) and a placebo (two initialisations of the same channel-independent configuration). It reported $+114$ (p=0.010) on Electricity, the registered primary, and $+318$ (p=0.001) on Traffic, with both controls null. We withdraw it. Three independent failures, reproduced by `scripts/e24b_reanalysis.py`:

The cancellation argument is false. Each arm's δ is normalized by its *own* clean error, so with $A(w) = \mathrm{MSE}(w) - m$ the contrast is $100(A_{\text{CD}}/m_{\text{CD}} - A_{\text{CI}}/m_{\text{CI}})$, which does not cancel when $m_{\text{CD}} \neq m_{\text{CI}}$ — and on Electricity m_{CD} exceeds m_{CI} by 30%. On a common denominator the registered primary falls to $+56$ with p=0.19.

The unit is wrong. The dose varies across windows, and all five seeds share the same ten of them, so window-level sampling error is common to every seed and does not shrink as seeds are added; a test over per-seed coefficients estimates training-run variation instead. At the window level the registered primary reads p=0.27 and Traffic p=0.07. Dropping any one window moves the Electricity coefficient between $+65$ and $+247$.

The controls could not have caught it. Widening or reseeding the channel-independent arm leaves its fragility profile essentially unchanged, so both contrasts vary by about 0.3%p across windows against the primary's 5.4–7.0%p. A control that never enters the regime where the confound lives is not evidence that the confound is absent.

This is the error our first attempt died of, moved one level down, and we committed it in the same days we were writing that diagnosis up. The stopping rule registered with the second test applies: no third estimator, and the coupling attribution is confined to the controlled data where it is §6's 11 of 12 configurations.

H CPiRi is permutation-equivariant by construction

Referenced from §6. CPiRi's spatial module is an `nn.TransformerEncoder` over channel tokens carrying no positional or index parameter on that axis, so the map commutes with a channel permutation. On the released 142.5M-parameter model (14.2M trainable, the rest a frozen encoder), permuting the input channels and permuting the module's representation back agrees to 9.5×10^{-7}. End to end the difference is 0.49 against a flow-sampling noise floor of 0.48 measured by running the same input twice — a floor that has to be measured before any such claim, since the model's head is stochastic. `scripts/e26_cpiri.py` reproduces both.

The five limits on the CPiRi contrast, in full. **(i)** The reported values are raw degradations, not control-subtracted excess, so they do not compare in magnitude to the $+3.2$%p paired flip. **(ii)** The spatial module trained on a fixed six-epoch budget with no validation selection, the setting §7 says inflates magnitudes. **(iii)** All 5 configurations are directed-lag, the topology where our own sweep finds the effect, so p=0.031 is conditional on that scope. **(iv)** Configurations four and five were added after the first three came back positive — at three the exact test floors at 0.125, and five is the smallest n reaching 0.05 — so the sequence is data-dependent and p=0.031 is nominal rather than exact. **(v)** Two configurations differ only in τ' and share a permutation arm, leaving four distinct relabeling measurements behind five signs; the shared value is the middle one.

I THE SEMI-REAL CHECK

Referenced from §6. We keep it out of the main text for the two reasons it states: the shift makes the task easier for every arm, and the injected coupling goes largely unused on real marginals, so there is little cross-channel reliance for it to disturb.

The semi-real check is weaker than it looks. On Weather marginals with injected coupling, Crossformer shows a positive excess under permutation ($+8.5\%$) while the equivariant models sit at ≈ 0, which matches the audit's ordering. We flag two reasons not to lean on it. First, the shift makes the task *easier* for every model—all arms improve, the CI control by 53% and Crossformer by 45%—so the excess is a difference between two large negatives rather than a degradation, and we exclude such shifts as stressors elsewhere (§3). Second, the injected coupling goes largely unused on real marginals, so this setting has little cross-channel reliance to disturb. We report the ordering as consistent with the audit, not as independent confirmation of it.

J THE THIRD ATTEMPT AT REAL-DATA ATTRIBUTION

`PREREG_E29.md` registers this before it ran, including that it is a third look and that its threshold is therefore $0.05/3 = 0.0167$. It uses no new runs: the degradations are E25's committed windows.

The design targets the two causes of the earlier failures. E23's coupling dose tracked general distribution movement, so here the cross-channel displacement of each evaluation window is regressed on its marginal displacement and only the residual is used; E24's cancellation failed on own-denominator terms at the seed unit, so the outcome is the common-denominator excess at the window unit.

It failed. On Electricity the single identity-storing arm that trains reads a slope of $+10.5$ at p=0.59; on Traffic, where every arm trains, nothing reaches the threshold and the equivariant arm's slope is the largest in magnitude and negative. The placebo — the channel-independent twin, which differs from the control only in seed — is null on both, which is the one thing that went as registered.

A deviation we have to declare. The rule as written returned a pass, carried entirely by Crossformer and TSMixer on Electricity: the two arms §6 excludes because they do not train under the shared recipe, with clean errors $30\times$ and $4.8\times$ the control's. The registration omitted that gate and we applied it after seeing the result. The gate is not new to this test — it is the one §6 already applies to those two arms, and E28 shows their failure is not a recipe artifact — but its own origin is post hoc too: it was added after E25's first run, when the rule as first written selected the most broken arm and returned a pass on a degenerate configuration. Applying it here is therefore a deviation twice over, and both readings are in the registration rather than only the one we prefer.

What the negative result does establish. On Traffic the marginal and cross-channel displacements correlate at $+0.96$, leaving the residual only 26% of the cross-channel spread; on Electricity they correlate at -0.16 and the residual keeps 99%. On the dataset the paper's ordering rests on, the two kinds of movement are close to inseparable. That is a reason this question is hard on real data rather than merely unanswered, and it is why we are not designing a fourth attempt.

K EVERY RESULT OVER A COMMON DENOMINATOR

Excess fragility divides each arm's degradation by that arm's own clean error, which is the failure mode §7 warns others about and which accounts for 22% of E27's interaction. The same question applies to every result the metric carries, so `scripts/denominator_sweep.py` recomputes them with both arms' absolute degradation scored over the channel-independent control's clean error.

Both orderings survive. The pre-registered audit's fragile arms still exceed its robust ones, by $+10.5\%$p; on Traffic the classes still separate, by $+11.0\%$p against 15.2%p at the own-denominator reading. The artifact runs in both directions — it inflates arms whose clean error is below the control's and deflates arms above it — but not symmetrically here: the control is the worst-fitting arm for five of the seven, and the three identity-storing arms the audit rests on are all on the inflated side,

so saying the direction cancels would be wrong. Recomputing is the answer to that rather than an argument about direction, and no arm the paper relies on changes sign. The two Electricity arms whose clean error is several times the control's do change sign, which is what their exclusion in §6 already says about them.

L E25 AT THE WINDOW UNIT

`scripts/e25_window_unit.py` recomputes §6's intervals with the evaluation window rather than the seed as the unit, from the same committed logs. On Traffic every identity-storing interval still excludes zero, with identity-storing lower bounds from $+3.1$%p, and the classes are still disjoint by $+2.7$%p against a same-class floor of 0.87%p at this unit. That floor is half the twin's window-unit interval, not the seed SD §6 uses: at the window unit there is one seed set and no seed spread exists, so the definition differs by construction and we say so rather than reuse the name. What does not survive is narrower and we withdraw it: the equivariant arm's interval becomes $[-15.5, +0.3]$, so beating the control outright is not established, only "less fragile than the identity-storing arms". On Electricity the single arm that trained reads $[-1.8, +8.1]$, which includes zero, so that dataset agrees in sign and does not on its own establish the effect. Adding seeds narrows a seed-unit interval by sampling training-run variation, which is not the variation an out-of-sample claim is about; the ten windows are also contiguous, so even these intervals are optimistic.

M IS THE ELECTRICITY EXCLUSION OUR RECIPE?

The exclusion in §6 removes the one arm that would contradict the audit, so it deserves more than an argument. `scripts/e28_recipe_sweep.py` fixes the rule and the consequences before running: an arm trains if its best-validation clean error lands inside the same $[0.5, 2.0]$ gate E25 applies, the sweep is learning rate $\in \{10^{-4}, 3{\times}10^{-4}, 10^{-3}\}$ against batch $\in \{8, 16\}$, and if any arm trains we report its excess fragility at its own optimum — including the outcome where Crossformer trains, still scores near zero, and becomes evidence against our own prediction, which we committed to putting in the body rather than a footnote.

Neither arm reaches the gate at any of the 6 settings. Crossformer sits between 29.6 and $29.9\times$ the control's clean error across all six — a spread of 0.3 while the learning rate moves by a factor of ten, and it trains for up to 29 epochs at the lowest rate rather than stopping early. TSMixer sits between 3.7 and $5.1\times$ and stops within four epochs everywhere. An optimisation failure is sensitive to the optimiser; this is not, which points at the architectures at this channel count rather than at the recipe. That is the outcome favourable to us, and it is reported here with the rule that would have produced the unfavourable one.

N HOW WIDELY IS THE SHUFFLING TEST RELIED ON?

An earlier version of this paper asserted, in the abstract and the introduction and without citation, that channel shuffling is *the field's* standard robustness diagnostic. A reviewer asked for the denominator, which is fair: the size of a corrective contribution is the extent of what it corrects. We looked, and report what we found rather than the sentence we started with.

What we can establish is narrow. Xu et al. (2026) runs the test, reports large degradations for channel-dependent baselines under it, varies its intensity through a settable shuffled-channel fraction, and builds a training objective and a deployment claim on it — that is one paper doing all four. Beyond it we did not find a body of work that presents channel-permutation robustness as evidence of structural robustness. A 2025 survey of channel strategies in multivariate forecasting (Qiu et al., 2026) organises the literature by channel treatment and does not carry permutation robustness as an evaluation axis at all, which is evidence against the practice being standard rather than for it.

We therefore state the premise as what it is. The claim this paper corrects is made explicitly by one contemporary paper, whose released code and weights we run directly (§6); whether others rely on the same inference implicitly is something we have not measured and do not assert. A reader who thinks the practice is rare should read our contribution as a correction to that paper and to the reasoning it exemplifies, not as a field-wide diagnosis.

O WHAT EACH SHIFT MODE DOES, PER TOPOLOGY

The names in the code and the text are the same, and `rewire` is the one that misleads: it is a *relabeling* in every topology, the axis a shuffling test covers, while the structural change the paper is about is `relag` on the chain and `reload` on the factor model. Table 8 is what a re-implementer needs and what Appendix C refers to when it says the factor *permutation* sweep was not run. The dose knob (`perm_swaps`) makes `rewire` exactly k transpositions from the identity in every topology.

Table 8: What each shift family does in each topology; `rewire` is the permutation a shuffling test covers.

topology	`rewire`	structural change	`sever`
diffusion	permute the adjacency's channel labels	—	coupling to zero
directed-lag	permute the chain order	`relag`: $\tau \to \tau'$, one lag shared by all parents	coupling to zero
latent-factor	permute the loading rows	`reload`: redraw the loadings	coupling to zero

P TWO LEMMAS, CHECKED ON THE GENERATOR

`scripts/theory_lyapunov.py` checks both on the generator and runs under `--verify`; the gate fails if either stops holding. Neither is a contribution: Lemma 2 is the definition of equivariance composed with a permutation-invariant loss, and Lemma 1's τ-invariance is a property of the chain's construction (a follower's own dynamics do not depend on τ), not a discovery. Both are stated because the arms' difference rests on their being exact, and checked because our first empirical check got one of them wrong.

Lemma 1 (a relabeling moves single-channel laws). Let x_t be a zero-mean stationary Gaussian linear process (the generator has no intercepts; with means, add equal means on each orbit to the condition below) with autocovariance sequence $\Gamma(l)$ and let P relabel the channels. The relabeled process has $\Gamma'(l) = P\Gamma(l)P^\top$ at every lag, so channel c receives channel $P^{-1}c$'s entire autocovariance sequence. The single-channel laws are preserved iff, on every orbit of P, the channels share one autocovariance sequence; a Gaussian channel's law is that sequence. Equal variances alone are necessary and not sufficient, and exchangeability of the joint law is sufficient and not necessary; an earlier statement conflated the two. On this chain the variances already differ, so no orbit of a non-identity relabeling preserves them. The lemma is about this chain, not a general impossibility: with $x_1(t) = \varepsilon_1(t)$ and $x_2(t) = \rho\,\varepsilon_1(t-\tau) + \sqrt{1-\rho^2}\,\varepsilon_2(t)$ for independent standard white noises, both channels are standard white noise, so swapping them preserves every single-channel law exactly while reversing which channel leads. A relabeling moves marginals only when the channels it swaps differ in law, as ours do. *Checked exactly on the chain* (block-Toeplitz Lyapunov, the method of Appendix D), whose variance falls $1.8\times$ from leader to last follower. For the permutations E27 drew, $\mathrm{diag}(\Sigma_{\text{post}})$ equals the permuted pre-break diagonal to 10^{-16}, a per-channel move of 0.010 (k=1) to 0.18 (k=8). Under a lag change every channel's own autocovariance function is invariant to τ at every lag (10^{-15}) while the cross-covariances move by up to 0.69 in standardised units: a lag change is a coupling change with every single-channel law fixed. Empirically (three seeds against a 20-seed floor of 0.032 ± 0.024) the relabeling's multiset of variances sits inside the floor (≤ 0.047) while its per-channel move reaches 0.19, and the lag change's per-channel move (≤ 0.049) sits inside the floor. A first version of this check, empirical only and on three seeds, read a three-seed difference as the lag change moving the multiset; over the same 20 seeds the lag change's move is 0.031 ± 0.023 against the floor's 0.032 ± 0.024, the three-seed gap was 1.1 standard errors, the exact computation shows it moves nothing, and the gate now reads the exact quantities.

Why the lag change holds every single-channel law fixed, and when it does not. The chain is $z_c(t) = a\,z_c(t-1) + b\,z_{p(c)}(t-\tau) + \varepsilon_c(t)$ with $|a| + |b| < 1$, independent white innovations, an autonomous leader and each follower with exactly one parent earlier in the order. Taking spectra, $S_c(\omega) = |1 - ae^{-i\omega}|^{-2}\big(b^2 S_{p(c)}(\omega) + \sigma^2\big)$: the lag enters only through $e^{-i\omega\tau}$, whose modulus is

one, so given the parent's spectrum the follower's does not depend on τ, and by induction from the leader neither does any channel's autocovariance sequence. The cross-spectrum between a follower and its parent carries $e^{-i\omega\tau}$ and moves. The argument needs a single parent: with several, the follower is driven by the parents' *average*, whose spectrum contains the parents' cross-spectra, and those carry lag phases that do not cancel. Measured directly on the generator at N=200,000, a lag change $4 \to 12$ moves per-channel variance by at most 1.7% with one parent, inside the sampling floor, and by 15.3% with eight. Every headline experiment uses one parent; the dense-coupling arm of Appendix E does not, and its lag change is therefore not a pure coupling change. Stationarity after the break is supplied by the evaluation gap of 800 points against a slowest mode of r=0.97, a relaxation time near 33 steps. `theory_lyapunov.py` confirms the implementation, not the claim.

Lemma 2 (exact zero for an equivariant model). If $f(Px) = Pf(x)$, the loss is permutation-invariant, the post-break law is the pre-break law relabeled, and the control and every step of the evaluation pipeline (normalisation, windowing) are themselves equivariant, then the population risk is identical for every P and the relabeling excess is identically zero. The equality is of population risks; two finite samples drawn independently differ by sampling error, and a pipeline that fits per-channel statistics on the unrelabeled series is not equivariant, which is where the residual below comes from. Checked with a trivially equivariant predictor (shared AR(1) plus the cross-channel mean): excess 0 to machine precision on relabeled data, and $+0.17\%$ once the same predictor is scored through per-channel standardisation fitted on the unrelabeled series. The residual is the pipeline's.

The trained models, numerically. Lemma 2 applies to a model only if it is equivariant as trained, which for the model the headline rests on had not been checked (CPiRi had, Appendix H). On 256 clean windows and 20 random permutations, iTransformer trained as in E27 satisfies $f(Px) = Pf(x)$ to $1.7e-7$ relative, PatchTST to $0.0e+00$, and Crossformer, which stores identity, breaks it at $9.6e-1$ (`scripts/check_equivariance.py`, under `--verify`).

Q When a frozen channel-dependent predictor loses to ignoring the other channels

This appendix states what excess fragility can and cannot say about the value of cross-channel information after a change, using a standard squared-loss decomposition that we do not claim as new. Let A be a target channel's own history and Y its future. When the change preserves the single-channel law, the best own-channel predictor $m_A = \mathbb{E}[Y \mid A]$ is the same before (P) and after (Q). Write the best all-channel predictors as $m_P = m_A + s_P$ and $m_Q = m_A + s_Q$. With every expectation under the *target* law,

$$R_Q(m_P) - R_Q(m_A) = \mathbb{E}_Q[s_P^2] - 2\,\mathbb{E}_Q[s_P s_Q], \qquad R_Q(m_A) - R_Q(m_Q) = \mathbb{E}_Q[s_Q^2],$$

so the frozen source predictor still beats the own-channel one after the change iff $2\,\mathbb{E}_Q[s_P s_Q] > \mathbb{E}_Q[s_P^2]$: it depends on how well the old cross-channel correction aligns with the new one, measured on the new law, and not on how much cross-channel value there is.

An exact construction shows both sides at once. Three channels mix two latent AR(1) factors with channel-specific lags, and only the lag matrix changes; every single-channel autocovariance is unchanged to machine precision while the input covariance moves by 0.51 in relative Frobenius norm. Before the change the all-channel optimum improves on the own-channel one, 0.955 against 1.184. After it, the frozen source optimum scores 1.294, worse than the own-channel 1.184 by 0.110, while the all-channel optimum for the new law scores 0.913: the cross-channel value is still there, 0.271, it has moved. So a positive excess fragility is evidence that a model relied on a relation that changed, not that cross-channel modelling was worthless after the change, and neither the ceiling nor the excess alone predicts which way a frozen model falls. `scripts/r3_risk_decomposition.py` computes every number here exactly and checks the identity to 10^{-12} under `--verify`; nothing is trained.

R A MARGINAL-PRESERVING CONSTRUCTION (E32)

Registered before the propositions were checked exactly, E32 was to rescale the per-channel innovation variances after a break so that every per-channel variance returned to its pre-break value, and to re-run E27's grid on that manifold (`PREREG_E32.md`). Implemented as one exact linear solve (`marginal_preserving=True`; the marginal variances are linear in the innovation variances), it settled itself before a model was trained. Under a lag change the projection is the identity, because every single-channel law of the chain is already τ-invariant (Lemma 1), so relag-mp *is* relag, and the run reproduces E27's relag cells to $0.0e+00$%p. Under a relabeling it is infeasible in 10 of twelve registered cells: the chain's variance falls monotonically downstream, and a relabeling that puts a high-variance channel upstream of a low-variance one needs a negative innovation variance to hold the downstream marginal. Gate G3 discards those cells, the registered interaction cannot be run, and by the registration no second construction is tried, so E32 is void by G3 on the relabeling side; the 2 feasible relabelings (both at k=1, two seeds) read $+3.9$ and -1.2%p, differ from their unprojected values by at most 0.17%p, and are reported without inference, with the projection moving their autocovariance dose by at most 3.4%. What the experiment leaves behind is sharper than what it was to test: on this generator the lag change already *is* a coupling change with every marginal fixed, and no relabeling of these channels can be made one.

S THE CAUSAL VERSION OF THE AUDIT

The audit of §6 is correlational: we read parameters, predicted, and observed. We also pre-registered the intervention — inject a per-channel identity embedding into iTransformer (predicted: becomes fragile) and replace Crossformer's per-channel encoder embedding with a shared one (predicted: becomes robust) — and it does not confirm the mechanism. Injection moves iTransformer the predicted way ($-1.6\% \to +8.8\%$, means over the $three$ seeds in the committed log) with an interval spanning zero; removal moves Crossformer the *wrong* way ($+20.9\% \to +25.7\%$). The likeliest reading is that our intervention was incomplete — we replaced only the encoder positional embedding, while Crossformer's decoder query slots and two-stage segment attention also index channels — but that is a hypothesis, not a result.

T IS THE AUDIT ARCHITECTURAL OR PIPELINE? E30 AND E31

E30 (void). Registered to test whether per-channel standardisation fitted before the break is what breaks iTransformer's equivariance as deployed. Under *frozen* statistics (E27's) its degradation at k=1, 2, 4, 8 is $+2.24$, $+3.14$, $+2.95$, $+4.15$%; with each window's own statistics it is $+2.21$, $+2.29$, $+2.19$, $+2.06$%. The design is void by its own second falsifier: the channel-independent control moved with it, because PatchTST is standardised by the same outer transform and so shared the mechanism. What survives is the observation, not a mechanism.

E31. The contrast E30 lacked is a class that stores identity in a parameter, which no input transform reaches. One model per (arm, seed), 12 trainings, scored on all eight (level, k) cells; a fresh PatchTST reproduces E30's per-seed values exactly ($+3.80$%, $+2.60$% at k=8). Own degradation at k=8 (seed SD up to 11.9%p at three seeds), frozen then adaptive: Crossformer $+42.1$/$+40.4$, TSMixer $+49.9$/$+48.6$, TimeXer $+41.9$/$+38.9$%; excess over E30's control under frozen $+38.3$, $+46.1$, $+38.1$%p, all above E2's $+10$%p FRAGILE threshold. Adaptive slopes on the autocovariance dose: $+81$ (p=0.025), $+91$ (p=0.021), $+73$ (p=0.057), four points each. The registration's verdict rule returns "any mix" for TimeXer's p=0.057, so the class claim is replaced by per-architecture reporting; no arm's fragility vanishes, which is the outcome that would have retracted the audit. E27, E30, E31 and E32 share seeds 4100–4102 and hence the same three data realisations: they are one generator's three draws re-scored under different conditions, not independent replications.

U THE AUDIT ON SIX ARCHITECTURES IT HAD NOT SEEN (E33): FALSIFIED

Registered in `PREREG_E33.md` and committed before the run, with the source-level audit of each architecture written down first. The rule, the thresholds and the training gate are E2's, unchanged;

the arms are TiDE, FreTS, MICN, Autoformer, FEDformer and SegRNN against E2's PatchTST control, five seeds, both break modes. No arm failed the training gate (clean errors 0.89–$1.02\times$ the control's). The seeds are E2's, so these arms see the same five data realisations E2's did; that is what "identical to E2" requires, and it is another reason the two sets are not pooled.

Excess on the permutation axis, five seeds: TiDE $+0.7$, FreTS -4.4, MICN $+6.7$, Autoformer $+22.7$, FEDformer $+31.3$, SegRNN -0.1%p. Three of the six — FreTS, MICN, SegRNN — fall on the wrong side of their registered threshold, which is the registration's falsifier 1: the audit does not generalise, Table 1's row stays *suggestive* at E2's 5 prospective arms, and the failures are reported by name and *not* pooled. Over the 11 prospective arms together the exact rank-sum probability is 0.0121 and the perfect separation E2 had is lost; the registration forbids using either number once falsifier 1 fires, and we do not. They are printed so that a reader who recomputes them does not find them missing.

The three misses are not one kind. MICN is short of the $+10$%p bar in the predicted direction with an interval covering it — E2's ambiguous band, scored as a miss because this registration has no such band. FreTS missed the ROBUST band *downward*, degrading less than the control, with an interval covering zero. SegRNN is the substantive failure: it holds a per-channel parameter, as the source audit said, and shows no fragility at all. Held to the stricter standard §6 applies to E2 — the whole 95% interval inside the predicted region, not just the point estimate — 2 of 6 arms are decisive here, TiDE and FEDformer; Autoformer's interval $[+5.2, +40.3]$ straddles the bar it clears on the point estimate.

Why, as a post-hoc diagnosis. We measured two architectural properties of each arm on an *untrained* model (Table 9) (`scripts/e33_mechanism.py`): whether output channel j moves when a different input channel is perturbed, and whether permuting the input permutes the output.

Table 9: Two properties of each E33 arm, measured on an untrained model, beside its permutation excess and the audit's call.

arm	mixes channels	equivariant	excess (%p)	audit
PatchTST (control)	no	yes	—	—
TiDE	no	yes	$+0.7$	hit
FreTS	no	yes	-4.4	miss
MICN	yes	no	$+6.7$	miss
Autoformer	yes	no	$+22.7$	hit
FEDformer	yes	no	$+31.3$	hit
SegRNN	no	no	-0.1	miss

The registered rule reads the equivariance column alone. SegRNN breaks equivariance and never mixes channels — its recurrent encoder runs per (batch, channel) and the per-channel embedding conditions only that channel's own decoder query — so a relabeling mismatches each channel with its embedding, and here that cost nothing. It need not in general: a channel-wise model with per-channel parameters fitted to channels with different dynamics would be hurt by a relabeling without ever mixing channels. So SegRNN's null is an observation about this generator, where the chain's channels differ mainly in scale, not evidence that mixing is necessary. Scored at the same thresholds, that conjunction calls 4 of 6 correct against the registered rule's 3. It was written after the outcome and refit on the very arms that falsified its predecessor, so it is a hypothesis and is not promoted into any claim in this paper.

Two arms ran with a longer decoder prefix than their papers use. The shared configuration passes `label_len=0`. Autoformer and FEDformer are the only arms here that slice with it, as `trend_init[:, -label_len:, :]`, and in Python `-0:` is `0:`, so both decoders received the *whole* lookback as their label prefix rather than the 48 steps their papers specify. They are also the two arms with the largest excess in the table above, so it is worth saying plainly: they trained normally, with clean errors inside the gate, and the prefix length does not touch the property the audit predicts from, which is whether a parameter is indexed by channel position. But the setting is not the one those architectures are reported at, and we did not notice until after the run.

A registered caveat we never exercised. The registration flagged FreTS's FFT along the channel axis as a mechanism a parameter audit cannot see. That branch is guarded by a comparison against the *string* `'0'`; E2's configuration, inherited unchanged, passes the integer `1`, so the arm we trained never ran it. Forced on, the branch moves the output by about $1e-4$ of its own scale at initialisation, which settles nothing about a trained model, and the registration allows no second design. The caveat stands untested and we report it as untested.

V THREE REGISTERED CHECKS THAT COULD NOT MOVE THE HEADLINE

Registered together (`PREREG_H3H5H6.md`) before any was computed; each runs under `--verify`.

H3: is the attribution failure predictable from the data? Not shown. E29's per-window marginal and cross-channel displacements are correlated $+0.96$ on Traffic and $+0.96$ on Solar but -0.16 on Electricity and -0.82 on ETTm2 (median $|r|$ 0.82 over 9 datasets). The registered proxy for coupling strength, mean $|\mathrm{corr}|$ between channels, does not order them: Spearman $+0.17$, exact p=0.68. Coupling strength by this proxy does not predict entanglement, and the nine values are description only.

H5: what fraction of the ceiling's fragility does the model pay? The generator is linear-Gaussian, so the optimal predictor fitted before the break and scored after it has an exact excess (Appendix D's method, on the directed-lag chain; the computation reproduces Lemma 1 to 10^{-14}). Under a relabeling it rises from $+7$ to $+52$%p across k while iTransformer's stays under $+0.4$%p (slope 0.002, p=0.970, seed-cluster t=0.1; not proportional). Under a lag change the two are proportional by the registered test — intercept -3.0 [-7.9, $+1.9$], slope 0.30 [0.18, 0.43], p=0.0004, seed-cluster t=11.0 — but the per-dose ratio rises from 0.11 to 0.27, so "fixed fraction" holds only at the resolution of a twelve-point intercept and we do not claim it. One registered gate, every exact variance within 3% of the generator's empirical one, failed at 3.8%: the tolerance was set below the sampling noise (exact SE 1.4%; largest deviation 2.2 standard errors) and is applied at three standard errors, a bound chosen after seeing that number, unregistered and favourable to us; disclosed.

H6: does the ceiling work as a switch? Not shown. On E25's ten Traffic windows the CD oracle at its capped depth never beats the CI oracle (clean -1.3%, windows -38.3 to -7.3%), so a switch keyed to the previous window's ceiling always chooses the channel-independent network and coincides with always-CI. That beats the two identity-storing arms always-CI already beats and loses to iTransformer (0.443 vs 0.457, CD better on 5 of ten windows) and TimeXer (0.452 vs 0.457). The clairvoyant switch fails the same way, so this ceiling does not track where CD networks fail on Traffic; at C=862 its four-lag cross-channel design against a ninety-six-lag channel-independent one is the handicap of Appendix D, in the regime where it dominates. The ceiling is a diagnostic here and does not work as a switch.

W THE TWO DISTURBANCES ON ONE DOSE PLANE

A reviewer put the objection this answers in its sharpest form: relag was run at three sizes and visibly scales with them, while `rewire` drew a single uniform permutation, so "a lag change hurts more than a relabeling" could be a statement about the sizes chosen rather than about the two families. Our existing defence — that the non-permutation moves marginals no more than the permutation does — answers a different objection: it rules out marginal movement, not magnitude in coupling space.

Design, registered before the runs. A permutation-intensity axis is not ours: Xu et al. (2026) already shuffles a settable fraction of channels (`NUN_SHUFFLENODES` in their runner), and varies it. What is missing is a scale on which that axis and a non-permutation axis can be compared, and that is what this adds. Our generator takes `perm_swaps=` k, building the permutation from k disjoint transpositions (k=1 one swap, k=$C/2$ a fixed-point-free involution), which gives a Cayley-distance dose rather than a fraction-of-channels dose. Both families are placed on a scale that is model-free and is what a forecaster depends on, $\|\Gamma_{\text{post}}-\Gamma_{\text{pre}}\|_F/\|\Gamma_{\text{pre}}\|_F$ over cross-channel autocovariances to

lag 15 on standardised channels, computed from each run's own series. Directed-lag, ring, C=16, ρ=0.6; `perm_swaps` $\in \{1, 2, 4, 8\}$ against $\tau : 4 \to \{5, 6, 8, 12\}$; three seeds, fixed in advance and not extended. The two families overlap in dose, which is what makes the comparison possible.

The registered test and its falsifiers. Pooled OLS excess $\sim$ dose + family + dose : family, with the prediction that the interaction is positive. It would have been falsified if the interaction were ≤ 0 or its interval covered zero with no matched-dose difference; if the permutation's response met or exceeded relag's; or if neither family responded to dose. The first outcome would have made us restate the headline as dose-conditional. It held: the interaction is $+38.6$%p $[+14.4, +62.8]$, p=0.003.

The two dose-responses. Over 24 runs the permutation's slope, -0.2%p per unit dose $[-12.7, +12.3]$, does not exclude zero while relag's does, $+38.4$%p $[+14.5, +62.4]$; the interaction is $+38.6$%p $[+14.4, +62.8]$, p=0.003 naive; positive under every unit and denominator we checked, and at the unit that is actually independent the enumerated test returns p=0.25, the floor for three blocks.

Unit and denominator. Two corrections a referee is owed, neither of which changes the sign. First the unit: the 24 observations are not independent, because inside one (family, seed) the training window is entirely pre-break, so the same fitted model is scored at every level — 6 clusters, not 24 points. The dependence sits in the intercept rather than the slope, so clustering *tightens* the interaction rather than loosening it (p=0.0013 cluster-robust); a wild cluster bootstrap, the appropriate test at this many clusters, gives p=0.031 enumerated exactly over all 64 sign vectors — which is also this test's floor at six clusters, 0.031. Second the denominator: excess fragility divides each arm by its own clean error, which is the failure mode we warn others about in §7. Scoring both arms' absolute degradation over the control's clean error instead gives $+30.0$%p (p=0.020 naive, 0.0039 clustered), so 22% of the coefficient is the denominator and the rest is not. A referee then found the correction to our correction: since the generator repair the two families share one trained model per seed — the clean error is identical per model across both families and every dose within a seed — so the six (family, seed) clusters are three independent blocks, not six. With the seed as the cluster the same enumerated test returns p=0.25, which is its floor at G=3 ($2/2^3$); the cluster-robust t is 8.5 and the coefficient does not move. Three blocks cannot establish the interaction at the independent unit whatever they show, and an earlier version of this paragraph claimed $p \leq 0.250$ under every unit and denominator; that is withdrawn. The remedy is independent blocks, which Appendix Y pre-registers on six fresh seeds.

A design limit worth stating because it bounds what could be concluded: with four dose points per family, the enumerated sign-flip test we use, clustering at the dose point, floors at $1/2^4$ for a single family's slope and cannot reach 0.05 whatever the data does, so the interaction carries the registered claim. We later added two points per family (`PREREG_E27B`, registered separately and leaving E27's test untouched). An earlier version said this lowered that floor to $1/2^6$; it does not, since the floor is set by independent blocks and the new points come from the same three seeds. At 6 points each, under a naive six-point t test that shares those seeds across points, relag's slope is $+44.9$%p per unit dose $[+30.5, +59.3]$, p=0.0010, and the permutation's is $+1.6$%p $[+0.1, +3.1]$, p=0.0449 — small, and under that test not zero; an earlier version called it still null, which the reported test does not support. What the two slopes show is a ratio, not a null.

What the flat permutation slope is evidence of. A referee argued the flatness is forced: `rewire` permutes the coupling graph, an equivariant model commutes with that, so there is nothing for the dose to move. The logs say otherwise — the equivariant arm's own degradation under permutation runs 2.24% to 4.15%, rising with permutation size, against 3.12% to 20.20% under the lag change. It moves; it moves much less.

What it moves *with* is worth stating, because it sharpens rather than weakens the comparison. Its rise tracks per-channel marginal displacement, which our normalisation — fixed at pre-break statistics — would be sensitive to; we report the association and have not isolated that mechanism from finite-sample effects. What the measurement does establish is the comparison: the displacement is *larger* for the permutation than for the lag change at every dose but the smallest, where the two coincide (0.024–0.088 against 0.019–0.025 mean relative change in per-channel standard deviation).

Lemma 1 says the lag change moves no single-channel law at all, so the lag arm's range is this measurement's finite-sample floor rather than a displacement, and the permutation reaches $4.7\times$ it. The disturbance that moves marginals more is the one that hurts less, and the ordering cannot be a marginal effect. An earlier version read *the disturbance that moves marginals two to seven times more is the one that hurts less*, at every level. That is withdrawn: the blanket form held only on the contaminated run of Appendix X. What the referee is right about is narrower: the excess subtracts a control whose own response varies by seed, so a single seed's permutation excess is dominated by that offset rather than by dose.

What it does and does not settle. It settles that the ordering is not an artifact of the disturbance sizes we picked, on this generator. The permutation's flat response is the sharper half: an equivariant model is close to indifferent to *how far* a relabeling moved, which is what equivariance would lead one to expect and what a shuffling test therefore cannot detect. It does not extend the finding beyond linear-Gaussian directed-lag coupling, and it is one architecture.

X A GENERATOR DEFECT, AND WHAT RE-RUNNING FIXED

Found by an audit of this repository four days before submission, and reported here because it changed published numbers.

The defect. The dose-controlled relabeling draws a permutation that is exactly k transpositions from the identity. That draw took the same generator that produces the series, so it consumed extra random numbers: a configuration with the dose knob set and one without produced *different pre-break data at the same seed*. The lag-change family never sets the knob, so every dose-controlled permutation arm sat on a different realization from the lag arm it is compared against. This is the defect our own generator comment says was fixed for the shift family, reintroduced one level down when the dose knob was added, and the unit test that pins the invariant for shift families was not parametrised over the knob. It is now.

The committed logs carried the fingerprint: the control's clean error was constant within each family and differed between them, on a segment that is entirely pre-break and must be bit-identical. Measured directly, the two families' control saw pre-break errors 13% apart at every non-zero dose, and identical when the knob was unset.

What we did. Repaired the draw to a side stream seeded off the configuration, which leaves the permutation deterministic in the seed and consumes nothing shared, then re-ran the affected experiments at their registered points, seeds and arms: the dose plane, its extension, the marginal-preserving construction, the normalisation switch and the identity-storing sweep. No design changed and no sample size was extended. The contaminated runs are kept beside the repaired ones in the artifact (`results/logs/phase_*_contaminated`) so the comparison can be checked rather than taken on trust.

What moved. The registered dose interaction, which is the headline this experiment tests, is *larger* on correctly paired data: $+30.8\%$p at $p{=}0.0087$ became $+38.6\%$p at $p{=}0.0034$. The permutation family's own dose slope, which the paper reads as the equivariant model being close to indifferent to how far a relabeling moved, fell from $+7.6$ to -0.2%p per unit dose — the defect had been manufacturing the slope our argument wanted to be absent. The confound was working against us, not for us.

Two things moved the other way and we state them. The permutation's measured marginal displacement fell from 0.048–0.134 to 0.024–0.088, which is why the blanket form of the marginal comparison in Appendix W is withdrawn and the claim now excludes the smallest dose. And the equivariant arm's own degradation under window-fitted normalisation at $k{=}8$ moved from -0.24% to $+2.06\%$; at three seeds its standard deviation is 3.29, so neither value is distinguishable from zero, and what the repaired run shows is the cleaner statement: under that switch the arm's degradation stops tracking permutation size, which is what Lemma 2 predicts, rather than reaching a particular small number.

The identity-storing sweep's verdict is unchanged: two arms clear both registered criteria, one is ambiguous on the slope test, and no arm's fragility vanishes.

Y REGISTERED FOLLOW-UPS (E34–E37)

Y.1 THE DOSE INTERACTION ON SIX FRESH SEEDS (E34)

Registered in `PREREG_E34.md` on the day this was found and committed before the run. Appendix W's three seeds cannot reach 0.05 at the seed unit, and adding seeds to a finished experiment after seeing it is the extension this paper discloses elsewhere as a defect, so this is a separate replication: E27's eight registered points, arms and harness on seeds 4103–4108, six blocks E27 never used. The primary test is the enumerated wild-cluster sign-flip test on the `dose:family` coefficient with the seed as the cluster over the six new seeds alone (G=6, floor $2/2^6$); the registration also fixes an effect size of interest, $+10\%$p per unit dose, below which a significant coefficient is reported as present but small.

Outcome: confirmed. On the six new seeds the interaction is $+40.2\%$p per unit dose, cluster-robust t=4.0, and the enumerated test returns p=0.0312, its floor at G=6. The enumeration is exact over all 2^6 sign vectors, but the test's size is exact only under its assumptions — independent seed blocks, residuals from the restricted model symmetric within a block, the regression correctly specified — and a p at the floor means the observed pattern is the most extreme available, not that the evidence is overwhelming. The coefficient clears the registered effect size of $+10\%$p per unit dose. The slopes the same regression implies are $+44.6$ for the lag change and $+4.4$ for the relabeling, whose CR1 interval at t_{6-1} is $[-9.2, +18.0]$: consistent with zero and with a small positive response alike, so we claim the difference and not the zero. Scoring both arms over the control's clean error instead gives $+33.0$ at p=0.0312. Pooled with E27's three seeds (G=9, secondary) the coefficient is $+38.2$ at p=0.0039, again the floor. E27 at the seed unit is what three blocks can say; E34 is the confirmatory test, and Table 1 rests on it.

Y.2 BOTH AXES WITH EVERY SINGLE-CHANNEL LAW FIXED (E35)

Every headline experiment uses the single-parent chain, where the lag change holds single-channel laws only by the single-parent argument and a relabeling does not hold them at all. E35, registered in `PREREG_E35.md` and committed with its generator, contract tests and analysis script before any training, removes both limits. Eight channels mix three independent AR(1) factors, $u_i(t) = \sum_k a_{ik} z_k(t - d_{ik}) + \varepsilon_i(t)$, with every loading of equal magnitude and random sign, so every channel has the same autocovariance and neither a redraw of the lag matrix D nor a channel permutation changes any single-channel law; the cross-covariances move under both. All shift segments come from one draw of factors and noise, and an unchanged copy is scored as a null control. Before training, the exact Gaussian oracle put the cross-channel gain at 13.2–19.7% of the own-history optimum on all 10 seeds (registered floor 2%), and the frozen source optimum degrades by 28.7% under the lag redraw and 75.5% under the relabeling.

Outcome: not confirmed, and the arm did not use the information. The registered primary is iTransformer's excess under the lag redraw minus its excess under the relabeling, seed as the unit, enumerated sign-flip test. It is $+0.71\%$p at p=0.0098: the sign is the predicted one but the mean is below the registered $+1.0\%$p, so the prediction fails. The registered reading of that failure is fixed by falsifier 3, which fired: iTransformer's clean error is -2.8% *worse* than PatchTST's (p=0.021; 1.067 against 1.038, with the exact own-history and all-channel optima at 0.862 and 0.715). The registration read that as "does not use the information"; E37 (Appendix Y.4) measures it directly and corrects the reading: on this generator iTransformer does rely on the other channels, with a reliance share of 0.23, but gains nothing from them — the same network fed one channel at a time fits better (0.960 against 1.047). A model whose use of a relation does not pay cannot show what happens when that relation moves, so E35 neither extends nor refutes the equivariant arm's lag-change result beyond the chain; we count it for neither side. Its null excess, $+1.3\%$p (p=0.053), is as large as its relabel excess.

The two arms that did use the other channels here — Crossformer (0.928) and TSMixer (0.907), both below the control — both store channel identity, and both pay on both axes: lag redraw +14.5 and +13.6%p, relabeling +28.1 and +37.5%p, each at p=0.002, the floor over 2^{10} sign vectors, with null controls of +2.4 and +1.4%p ($p \geq 0.166$). Because this relabeling moves no single-channel law, their relabeling cost cannot come from changed marginals, which on the chain it could. That is the result E35 adds: on a dense multi-parent generator, a model that uses cross-channel structure pays when the structure changes with every marginal fixed, on either axis. The equivariant half of the contrast remains a single-parent-chain result. Every trained arm is far from the exact own-history optimum here (clean errors 0.907–1.067 against 0.862): the protocol's fixed training budget was not tuned for this generator, and the magnitudes should be read with that in mind.

Y.3 The same model with and without channel access (E36)

The body reads iTransformer's clean error below the control's as consistent with coupling use, not proof of it, because the two are different architectures. E36, registered in `PREREG_E36.md` before it ran, removes the confound: iTransformer-CI is the same class, configuration, initialisation and parameter count, applied to each channel separately, so its variate attention sees one token and nothing crosses channels (replacing the other channels changes its forecast for a channel by exactly zero; checked before training). E2's chain, 10 fresh seeds, the E2 harness. The primary is iTransformer's lag-change excess minus iTransformer-CI's, per seed, under the enumerated sign-flip test, with a registered effect size of +3.0%p.

Outcome: confirmed. Channel access accounts for +6.0%p of the lag-change excess (p=0.002, the floor; positive on 10 of 10 seeds): +7.1%p with access, +1.1%p without. Access also lowers clean error by 21.9% (p=0.002), so the equivariant arm does use the coupling it later pays for when the lag moves. Under the relabeling, for which nothing was registered, the access effect is +1.5%p (p=0.055), and iTransformer's own relabeling excess is +2.7%p (p=0.025): small and not zero, consistent with E34's relabeling slope and its interval. The result is one architecture on one generator.

Y.4 Reliance on the other channels, read without retraining (E37)

Excess fragility says what a model pays when a relation changes; it cannot say whether a model that pays nothing is robust while using the other channels or does not use them. E37 (`PREREG_E37.md`) measures the missing quantity on the same checkpoint. For a target channel j with own history A, other channels' history B and future Y, $U_f(j) = \mathbb{E}[(Y - f(A, B'))^2 - (Y - f(A, B))^2]$ with B' drawn exactly from $p(B \mid A)$, which both generators give in closed form; the reliance share divides $\sum_j U_f(j)/2$ by the exact available value $\sum_j (R^*_{\mathrm{CI}} - R^*_{\mathrm{CD}})$. This is conditional permutation importance for a group of channel histories, which we do not claim as new; the exact sampler is what the generators add. A model that reads only its own history scores zero sample by sample (here at most 1×10^{-16} for PatchTST and iTransformer-CI), the Bayes predictor scores one, and for any other model the share can leave $[0, 1]$. `timecard.depshift.reliance` releases it.

Outcome: confirmed, as an instrument check. Both registered predictions were expected from E35 and E36 and test whether the statistic reads, without retraining, what those ablations saw. iTransformer's share is 0.72 on the chain (p=0.002, registered floor 0.25) and 0.49 lower on the lagged factors (p=0.002), where it is 0.23. The two cells excess alone cannot separate are then separated. On the chain the model relies on the other channels heavily and pays little for a relabeling (+2.4%p, p=0.082) but pays for a lag change (+5.3%p, p=0.006): robust to one shift *while* using the coupling. On the lagged factors it relies on them less, fits worse than the same network without them, and pays for neither shift (+0.6 and +1.2%p). TSMixer relies on them in both families (0.91, 0.64) and pays for both shifts in both. Across seeds on the chain, iTransformer's share and its clean gain from access agree only loosely (rank correlation 0.42, ten seeds). The share is a property of a fixed model on the clean regime; it does not predict a cost by itself, and nothing here is on real data, where no exact sampler exists.

Z TWO NEGATIVE RESULTS OUTSIDE THE MAIN CLAIM

These arose while building the study and are reported for completeness; neither bears on the recoverability moderator, so they are kept out of the main text.

Counterfactual regularization is inert. Aiming to *reduce* the fragility we characterize, we tested a two-sided counterfactual regularizer whose sensitivity arm lower-bounds a model's response to dependency-violating perturbations. It is inert for normally-trained models: the hinge is satisfied by default—the response to a large perturbation already exceeds the margin—so the term contributes no gradient and leaves both accuracy and robustness unchanged. Rescaling the response by the perturbation norm does not fix this, because the objective asks for *more* sensitivity whereas robustness wants sensitivity *controlled*. Forcing sensitivity is not the same as controlling it.

An isometric front-end and its certificate. Where an isometric front-end is useful as a control, we use a learned lifting-wavelet decomposition that remains an exact isometry through training (operator norm 1.0), admitting a per-instance Lipschitz certificate that held with zero violations across 384 tests. We report it as a tool rather than a contribution of this paper.