A TREATMENT-ANCHORED PROXY DIAGNOSTIC FOR HIDDEN CONFOUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Can indirect measurements of hidden factors help detect confounding between a treatment and an outcome? AO-CES measures how much these proxies change the predicted treatment. It then tests whether that change predicts the outcome after accounting for treatment and recorded covariates. We characterize misleading proxy mixtures in linear Gaussian models and correct the leading estimation bias under explicit learning conditions. In a controlled Gaussian study with 3,200 observations, correction reduces false positives from 90% to 5.5% while retaining 100% detection. Across 21 benchmark conditions, full-proxy prediction is competitive with five feature-learning variants, including variational autoencoders. As an application after Fast Causal Inference, AO-CES detects 93–100% of confounded pairs in structured graphs. These results show how proxy measurements can strengthen the assessment of hidden-confounding risks and guide closer causal scrutiny of observational findings.

1 INTRODUCTION

Many observational studies ask whether changing a particular treatment T would change an outcome Y, rather than seek an entire causal graph. Examples include treatment access and health, job training and earnings, and interventions in social networks (Louizos et al., 2017; LaLonde, 1986; Veitch et al., 2019). An unmeasured common cause can make the observed association differ from the effect of intervening on T (Pearl, 2009; Peters et al., 2017). Negative controls and proxy methods use additional measurements to investigate or adjust for this problem (Lipsitch et al., 2010; Miao et al., 2018). We study a diagnostic question: do pretreatment measurements provide evidence of hidden confounding for a scientifically specified treatment–outcome relation?

Consider whether receiving a medication improves recovery: $T = 1$ denotes receiving it and $T = 0$ not receiving it; Y is a recovery score, for example physical functioning measured from 0 to 100. Following the treatment-access example of Louizos et al. (2017), household resources may affect both access to medication and recovery through living conditions and support. When these resources are not directly measured in the research data, they are a *hidden confounder*, denoted by U. Suppose a recorded estimate of pretreatment household income, W, measures those resources imperfectly. This indirect measurement is a *proxy*; it is informative about U without observing U itself.

For example, $\boldsymbol{C}_T = (\text{age}, \text{site})$ predicts the probability of medication receipt among patients of the same age and site. The predictors $\boldsymbol{C}_Y = (\text{age}, \text{site}, \text{comorbidities})$, together with T, predict recovery. We call these predictor sets the *contexts*: we adjust for them from the outset, whereas we test whether W adds information. One proxy W or a vector $\boldsymbol{W}$ can serve this role; proxies need not be independent of $\boldsymbol{C}_T, \boldsymbol{C}_Y$.

A structural causal model (SCM) specifies how each variable is generated from its causes (Pearl, 2009). An arrow $U \to T$ means that changing U can change T through its generating mechanism. The path $T \leftarrow U \to Y$ is a *back-door path*: it starts with an arrow into T and creates an association not caused by changing treatment. Assigning T externally breaks its link from resources but leaves their influence on recovery intact. We assume the causal ordering T before Y; we do not infer it or establish a nonzero treatment effect.

Two difficulties arise when looking for such a hidden cause. First, confounded and unconfounded models can have the same distribution of (T, Y) (Peters et al., 2017; Sanjaroonpouri & Ramazi, 2025).

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.

Second, proxies may contain several unrelated latent signals. In Figure 1(b), U_T affects treatment T but affects Y only through T; U_Y affects outcome Y but not T. A learned summary $Z = r(\boldsymbol{W})$ can mix their separate measurements $\boldsymbol{W} = (\boldsymbol{W}_T, \boldsymbol{W}_Y)$: one part predicts T, while another predicts Y after accounting for T. Thus the same summary can be associated with both variables without measuring a shared cause. Known measurement groups could be tested separately; with unknown roles, we must choose the summary carefully. Negative controls and latent-variable models already distinguish shared from treatment-only and outcome-only factors (Lipsitch et al., 2010; Zhang et al., 2021).

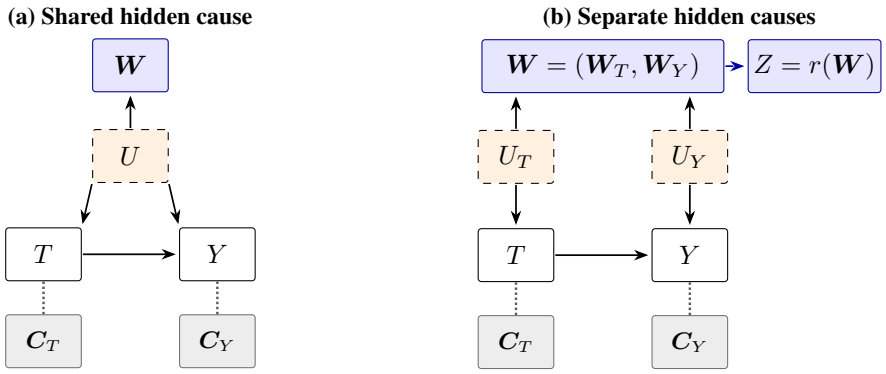

Figure 1: (a) $\boldsymbol{W}$ measures a shared hidden cause U. (b) Separate measurements form $\boldsymbol{W} = (\boldsymbol{W}_T, \boldsymbol{W}_Y)$; mixing them into Z can predict both T, Y without a shared cause. Solid arrows are causal; the blue dashed arrow computes Z. Dashed boxes are unobserved. Gray context boxes and dotted links denote adjustment, not causal relations; context-related causal arrows are suppressed.

The *Anchored Orthogonal Confounding-Evidence Screen* (AO-CES) uses the same proxy summary to test treatment relevance and adjusted-outcome association. It predicts treatment T twice: once using $\boldsymbol{C}_T$ alone and once adding $\boldsymbol{W}$. It subtracts the first prediction from the second. This difference is the *treatment anchor* A_0, a projection used in projected covariance testing (Lundborg et al., 2024). In our example, adding the income measurement might change a patient's predicted probability of treatment from .40 to .60, giving an anchor of .20. On separate test observations, AO-CES asks whether this score predicts T beyond $\boldsymbol{C}_T$, and whether it predicts Y beyond $(T, \boldsymbol{C}_Y)$. "Orthogonal" means that small regression errors do not change the test moment to first order, reducing fitting-induced associations (Section 4.2). Under Section 2.2's conditions, evidence for both supports reconsidering the measurements, adjustment, or study design; a negative result does not establish that confounding is absent.

The diagnostic has two separate requirements. First, measurements $\boldsymbol{W}$ must support an interpretation about confounding of $T \to Y$: a noisy record mixing separate causes of T and Y can mislead the test even if $\mathbb{E}[T \mid \boldsymbol{W}, \boldsymbol{C}_T]$ is known exactly. Second, errors in estimating that prediction can introduce outcome-only information into an otherwise valid anchor A_0.

1.1 Related work and the research gap

Negative controls test associations excluded by a scientific model (Lipsitch et al., 2010); future-treatment proxies offer one such hidden-bias test (Elwert & Pfeffer, 2022). Lee (2014) compare crude and proxy-adjusted risk ratios and already examine treatment-only and outcome-only factors. Proximal methods identify effects under measurement restrictions (Miao et al., 2018; Mastouri et al., 2021; Park et al., 2024); DANCE and related methods select proxies under restricted linear models (Kummerfeld et al., 2024; Xie et al., 2024). We instead use treatment prediction to select proxy information for a diagnostic.

Propensity scores summarize treatment assignment (Rosenbaum & Rubin, 1983); for binary T, our anchor is the difference of two such probabilities. The projected covariance measure (PCM) tests a learned projection's residual association with a response (Lundborg et al., 2024). The generalized covariance measure (GCM), its weighted extensions, and double machine learning provide residual-product tests and orthogonalization (Shah & Peters, 2020; Scheidegger et al., 2022; Chernozhukov

et al., 2018). AO-CES uses the information selected for predicting T in an additional comparison with adjusted Y. We correct its additional estimation error using locally robust estimation for generated regressors (Chernozhukov et al., 2022; Escanciano & Pérez-Izquierdo, 2026).

The proxy tests of Liu et al. (2024) and Wu et al. (2025) ask whether a causal relationship remains after accounting for a latent variable, not whether an existing relationship is also confounded. Higher-order least squares (HOLS) and kernel regression confounder detection (KRCD) diagnose confounding through regression comparisons under model restrictions (Schultheiss et al., 2024; Chen et al., 2026). Other methods exploit changes across environments (Karlsson & Krijthe, 2023; 2025; Mameche et al., 2024; Reddy & Balasubramanian, 2024) or elicit confounders from decision-makers (Plečko et al., 2026). Our information comes from proxies in one population (Appendix E.1). Graph-learning methods target a larger structure (Colombo et al., 2012; Maeda & Shimizu, 2020; 2021); methods for adding newly observed variables address a related issue: a previously hidden common cause may become available to explain an existing association (Kalantari et al., 2025).

The causal effect variational autoencoder (CEVAE) and multimodal models use noisy measurements for effect estimation (Louizos et al., 2017; Deshpande et al., 2022); TEDVAE separates treatment-only, outcome-only, and shared factors to estimate treatment effects, whereas we test for evidence of confounding (Zhang et al., 2021). Effect estimators also construct treatment-informed proxies or tolerate invalid proxies (Kottler & Cochran, 2026; Yu et al., 2025). Our diagnostic requires preservation of measurement restrictions, not just accurate prediction; model misspecification remains consequential (Rissanen & Marttinen, 2021).

1.2 Contributions

Our contributions distinguish measurement problems from prediction error: (1) in a linear Gaussian model, an exact expression showing when proxy measurements can create apparent confounding despite the absence of a hidden common cause **(Proposition 1)**; (2) a correction that removes the leading false association caused by prediction error, together with conditions under which the fitted test controls false positives **(Theorem 1, Corollary A.1)**; (3) matched experiments separating measurement quality, choice of predictor, and weak signal, including an explanation of limited binary-treatment detection **(Section 5)**; (4) an FCI application that adds proxy-based evidence of hidden common causes to discovered relationships, with false-annotation control and an evaluation of the trade-off between testing fewer candidates and missing confounded pairs **(Section 5.1, Lemma H.1)**.

2 Observed data and causal interpretation

2.1 Measurements and the question

For each of n independent, identically distributed records, we observe $O_i = (\boldsymbol{W}_i, \boldsymbol{C}_{T,i}, \boldsymbol{C}_{Y,i}, T_i, Y_i)$, where $i = 1, \ldots, n$. The treatment T may be binary or continuous, and T and Y have finite second moments. All context and proxy measurements precede treatment. The contexts and proxy measurements depend on the application; Table 1 collects the notation for reference.

The observed-data question has two parts: does the proxy score help predict T beyond $\boldsymbol{C}_T$, and does it help predict Y beyond $(T, \boldsymbol{C}_Y)$? Including T avoids mistaking treatment-mediated prediction for confounding. The causal interpretation requires the following restrictions on proxy information (Lipsitch et al., 2010; Elwert & Pfeffer, 2022).

2.2 Conditions for a confounding interpretation

The ideal score is the treatment-prediction difference

$$A_0(\boldsymbol{W}, \boldsymbol{C}_T) = \mathbb{E}[T \mid \boldsymbol{W}, \boldsymbol{C}_T] - \mathbb{E}[T \mid \boldsymbol{C}_T]. \tag{1}$$

This is the unweighted projection used by Lundborg et al. (2024, Sections 1.1–2), with their response, tested predictors, and conditioning variables relabeled as T, $\boldsymbol{W}$, and $\boldsymbol{C}_T$. Section 3.2 explains why it measures the proxy contribution. For a causal interpretation, we require the following two conditions. Here $\boldsymbol{U}$ includes all relevant unmeasured pretreatment factors, including separate factors when there is no shared cause.

Assumption 1 (Proxy mean exclusion)**.** *The designated pretreatment measurements $\boldsymbol{W}$ satisfy*

$$\mathbb{E}[Y \mid T, \boldsymbol{C}_Y, \boldsymbol{U}, \boldsymbol{W}] = \mathbb{E}[Y \mid T, \boldsymbol{C}_Y, \boldsymbol{U}]. \tag{2}$$

Once $T, \boldsymbol{C}_Y, \boldsymbol{U}$ are fixed, the measurements add no mean-outcome information. The income record measures household resources rather than independently changing recovery; an additional income-related influence would violate this interpretation.

Assumption 2 (Ideal-anchor validity under no hidden confounding)**.** *If there is no hidden common cause of T and Y—that is, no unobserved factor influences both T and, through a route bypassing T, Y—then*

$$\mathbb{E}[Y \mid T, \boldsymbol{C}_Y, A_0] = \mathbb{E}[Y \mid T, \boldsymbol{C}_Y]. \tag{3}$$

In the medication example, suppose no hidden common cause affects both receiving medication and recovery. The change in predicted receipt obtained by adding the income measurement must then give no extra information about mean recovery beyond T, age, site, and comorbidities. This requirement does not follow from pretreatment timing: a summary can still mix separate treatment and outcome information. Section 4.1 derives concrete measurement conditions and a counterexample; Section 4.2 addresses the additional issue created by estimating A_0.

Under these assumptions, an adjusted association between A_0 and Y contradicts what the no-confounding model predicts.

3 Constructing and testing the proxy score

We construct the score and its two comparisons; Section 4 supplies the fitting correction, and Algorithm 1 summarizes implementation.

3.1 Why use treatment prediction?

Testing whether an arbitrary summary predicts both treatment T and outcome Y is insufficient. For illustration, let U_T, U_Y be independent standard normal variables, $T = U_T$, $Y = \beta T + U_Y$, and $\boldsymbol{W} = (U_T, U_Y)$. Here β is the direct effect and there is no shared hidden cause. The summary $Z = U_T + U_Y$ predicts T through U_T and adjusted Y through U_Y. The treatment association is $\mathrm{Cov}(T, Z) = 1$. For the outcome comparison, subtract the predictions based on T: $Y - \beta T = U_Y$ and $Z - \mathbb{E}[Z \mid T] = U_Y$. Thus $\mathbb{E}[(Y - \beta T)\{Z - \mathbb{E}[Z \mid T]\}] = \mathbb{E}[U_Y^2] = 1$: different factors supply the two associations. Predicting treatment from the complete $\boldsymbol{W}$ instead selects only U_T here.

3.2 What the difference of expectations means

Write $m_0(\boldsymbol{C}_T) = \mathbb{E}[T \mid \boldsymbol{C}_T]$ for the context-only prediction and $m_1(\boldsymbol{W}, \boldsymbol{C}_T) = \mathbb{E}[T \mid \boldsymbol{W}, \boldsymbol{C}_T]$ for the prediction after adding the proxies. Equation (1) subtracts these predictions for the *same* record. In the running example, $.60 - .40 = .20$ means that the income measurement raises predicted treatment receipt by 20 percentage points. It is not a treatment effect or a comparison between two patients.

Equivalently, $A_0 = \mathbb{E}[T - m_0(\boldsymbol{C}_T) \mid \boldsymbol{W}, \boldsymbol{C}_T]$. The quantity $T - m_0(\boldsymbol{C}_T)$ is treatment variation left after recorded predictors are accounted for; A_0 is the part of that variation predictable from the proxies. The outcome comparison uses this same treatment-selected information.

The ideal prediction averages correctly within each context group, so

$$\mathbb{E}[A_0 \mid \boldsymbol{C}_T] = 0. \tag{4}$$

Among patients with the same age and site, proxy-augmented predictions average back to the context-only prediction. Individual changes can be positive or negative even though their conditional average is zero.

Centering makes the proxy contribution interpretable; prediction of T selects its information. In the running example, centering cancels from the linear outcome comparison, but treatment-based selection remains essential (derivations in Appendix A.2).

The expectations above describe an ideal population quantity; in data, we must estimate them. Use *development data*, observations reserved for fitting and tuning, to train $\widehat{m}_1$ on $(\boldsymbol{W}, \boldsymbol{C}_T)$ and $\widehat{m}_0$ on $\boldsymbol{C}_T$. Their difference estimates A_0:

$$\widehat{A}_0 = \widehat{m}_1(\boldsymbol{W}, \boldsymbol{C}_T) - \widehat{m}_0(\boldsymbol{C}_T). \tag{5}$$

We evaluate $\widehat{A}_0$ on separate *test data*, unused in fitting. An optional learned representation $\boldsymbol{Z} = r(\boldsymbol{W})$ adds predictors to $\widehat{m}_1$ without replacing $\boldsymbol{W}$ (Appendix G).

3.3 Two questions about the same score

Let A stand for the ideal score A_0 or its estimate $\widehat{A}_0$. Choose transformations $h_1, \ldots, h_J$ before testing: $h_1(a) = a$ checks signed co-movement, while squaring can reveal patterns involving large scores of either sign.

The *treatment test* removes the context prediction from both T and the transformed score:

$$R_T = T - \mathbb{E}[T \mid \boldsymbol{C}_T], \qquad Q_{T,j} = h_j(A) - \mathbb{E}[h_j(A) \mid \boldsymbol{C}_T], \qquad \theta_{T,j}(A) = \mathbb{E}[R_T Q_{T,j}]. \tag{6}$$

The residuals R_T and $Q_{T,j}$ are deviations from what is expected given $\boldsymbol{C}_T$. Their average product $\theta_{T,j}$ measures co-movement. With $h_1(a) = a$, Equation (4) gives $\mathbb{E}[h_1(A_0) \mid \boldsymbol{C}_T] = 0$. For a treated patient with context-only probability .40 and ideal anchor .20, $R_T = 1 - .40 = .60$ and $Q_{T,1} = .20 - 0 = .20$, giving product .12. Consistently nonzero average products indicate additional treatment information.

The *outcome test* instead removes the predictions based on $(T, \boldsymbol{C}_Y)$:

$$R_Y = Y - \mathbb{E}[Y \mid T, \boldsymbol{C}_Y], \qquad Q_{Y,j} = h_j(A) - \mathbb{E}[h_j(A) \mid T, \boldsymbol{C}_Y], \qquad \theta_{Y,j}(A) = \mathbb{E}[R_Y Q_{Y,j}]. \tag{7}$$

We now compare records with the same $(T, \boldsymbol{C}_Y)$. If recovery is five points above its adjusted prediction and the anchor is .10 above its adjusted prediction, their product is .50. Including T is essential: otherwise the score could predict recovery merely because it predicts receipt of an effective medication. In Figure 1(a), holding T fixed blocks the route $\boldsymbol{W} \leftarrow U \rightarrow T \rightarrow Y$, while $\boldsymbol{W} \leftarrow U \rightarrow Y$ can remain informative. The residual products measure the corresponding association in conditional means; they are not a general test of graphical separation. For the linear choice $h_1(a) = a$, $\theta_{T,1}(A) = \mathbb{E}[R_T A]$ and $\theta_{Y,1}(A) = \mathbb{E}[R_Y A]$, because the subtracted conditional predictions have zero average product with their respective residuals.

These GCM residual moments (Shah & Peters, 2020; Scheidegger et al., 2022) test whether all $\theta_{T,j} = 0$ or all $\theta_{Y,j} = 0$, not every form of conditional dependence. For the ideal anchor, the linear treatment moment satisfies

$$\theta_{T,1}(A_0) = \mathbb{E}[(T - m_0)A_0] = \mathbb{E}[A_0^2]. \tag{8}$$

To see why, write $T - m_0 = (T - m_1) + A_0$. The prediction error $T - m_1$ has conditional mean zero given $(\boldsymbol{W}, \boldsymbol{C}_T)$, so its product with A_0 averages to zero. A nonzero ideal anchor therefore has treatment relevance automatically. The held-out treatment test checks whether the *fitted* score retains that relevance.

Combining the answers. A random-sign multiplier bootstrap calibrates the treatment test (Appendix B.2) (Chernozhukov et al., 2013). The outcome test uses the corrected statistic and calibration derived in Section 4.

Let p_T and p_Y be the treatment and corrected outcome p-values. The reported value is

$$p_{\mathrm{AO}} = \max\{p_T, p_Y\}. \tag{9}$$

At level $\alpha = .05$, for example, $(p_T, p_Y) = (.01, .18)$ is insufficient: only treatment relevance is detected. The pair $(.01, .02)$ supports both associations because $p_{\mathrm{AO}} < \alpha$. Outcome prediction alone is also insufficient: an outcome-only factor could explain it. This intersection–union rule (Berger & Hsu, 1996) checks fitted-score treatment relevance but targets the ideal anchor in the corrected outcome test. Under its stated learning conditions, Theorem 1 bounds false positives when the *ideal* anchor A_0 has no adjusted-outcome association, even though the implemented test uses an estimate of that anchor.

4 MEASUREMENT CONDITIONS AND FITTED-ANCHOR VALIDITY

4.1 WHEN THE MEASUREMENTS SUPPORT THE INTERPRETATION

Raw measurements can already mix outcome-only information. Consider the linear Gaussian special case with constant contexts and continuous T and Y:

$$T = \boldsymbol{a}^\top \boldsymbol{U} + \varepsilon_T, \quad Y = \beta T + \boldsymbol{b}^\top \boldsymbol{U} + \varepsilon_Y, \quad \boldsymbol{W} = L\boldsymbol{U} + \boldsymbol{\varepsilon}_W.$$

Here $\boldsymbol{U} \sim N(0, I_p)$ contains p independent latent factors; $\boldsymbol{a}, \boldsymbol{b}$ are their treatment and outcome coefficients, β is the direct effect, and L maps factors to measurements. Assume mutually independent, centered Gaussian latent factors and error vectors, allowing correlated measurement errors. Define $D = \mathrm{Var}(T) > 0$ and $Q = \mathrm{Var}\{\mathbb{E}[\boldsymbol{U} \mid \boldsymbol{W}]\}$: its entries describe how the predictable parts of the latent factors vary together. For the identity transformation $h_1(a) = a$, recall that $\theta_{Y,1}(A_0) = \mathbb{E}[(Y - \mathbb{E}[Y \mid T])\{A_0 - \mathbb{E}[A_0 \mid T]\}]$ measures association after adjusting both outcome and anchor for treatment.

Proposition 1 (Measurement geometry)**.** *Under this model, with nonsingular* $\mathrm{Var}(\boldsymbol{W})$*,*

$$\theta_{Y,1}(A_0) = \boldsymbol{b}^\top Q \boldsymbol{a} - \frac{\boldsymbol{a}^\top \boldsymbol{b}}{D} \boldsymbol{a}^\top Q \boldsymbol{a}. \tag{10}$$

This moment is zero exactly when $\mathbb{E}[Y \mid T, A_0] = \mathbb{E}[Y \mid T]$.

Without a shared cause, $\boldsymbol{a}$ and $\boldsymbol{b}$ use disjoint factors, leaving $\boldsymbol{b}^\top Q \boldsymbol{a}$. Separate measurement channels with independent errors make this zero; an invertible transformation of the complete observed vector preserves that property. For $T = U_T + \varepsilon_T$ and $Y = \beta T + U_Y + \varepsilon_Y$, a single record $W = U_T + U_Y + \varepsilon_W$ mixes the independent causes and gives a positive moment even when the prediction function is known exactly. The analogous concern in the medication example is a record combining access resources and unrelated recovery support. Correcting fitting error cannot remove this population association. Appendix A.6 proves the result; Proposition A.1 bounds compression loss.

4.2 REMOVING THE LEADING FITTING ERROR

In the medication example, a fitted model may overpredict receipt for patients whose measurements predict unusually good adjusted recovery. Among these patients, observed minus predicted receipt is negative on average; the correction uses this error to offset the spurious outcome association. An independent test sample does not remove systematic fitting error.

Write $X = (\boldsymbol{W}, \boldsymbol{C}_T)$, $C = \boldsymbol{C}_T$, and $V = (T, \boldsymbol{C}_Y)$. Recall $m_1(X) = \mathbb{E}[T \mid X]$, $m_0(C) = \mathbb{E}[T \mid C]$, and $A_0 = m_1 - m_0$. For binary receipt, these conditional means are treatment probabilities. Write $g(V) = \mathbb{E}[Y \mid V]$ and $R_Y = Y - g(V)$ for adjusted recovery. For each chosen transformation h_j, let $r_j(V) = \mathbb{E}[h_j(A_0) \mid V]$ and define

$$q_j(X) = \mathbb{E}[(Y - g(V))h_j'(A_0) \mid X], \qquad k_j(C) = \mathbb{E}[(Y - g(V))h_j'(A_0) \mid C]. \tag{11}$$

Here h_j' is the derivative of h_j. For $h_1(a) = a$, q_1 predicts adjusted recovery using proxies and treatment context; k_1 uses context alone. The corrected product is

$$\psi_j = (Y - g(V))\{h_j(A_0) - r_j(V)\} + q_j(X)\{T - m_1(X)\} - k_j(C)\{T - m_0(C)\}. \tag{12}$$

Both added terms average to zero at the true predictions, preserving the target $\theta_{Y,j}(A_0)$ from Equation (7). With fitted predictions, they cancel the leading anchor error; remaining bias involves products of estimation errors. For the medication example, $\boldsymbol{C}_T \subseteq \boldsymbol{C}_Y$ gives $k_1 = 0$ at the population level, leaving the q_1 correction. Appendix A.5 gives an exact cancellation example; the construction applies locally robust estimation for generated regressors (Chernozhukov et al., 2022; Escanciano & Pérez-Izquierdo, 2026).

Split development data into two approximately equal parts, the *balanced inner folds*. Fit m_1, m_0, g on one part and predict the other, then swap. These *out-of-fold predictions* use models not trained on the predicted records and supply targets for r_j, q_j, k_j. Refit learned representations within each fold (Algorithm 1). Outcome data train the outcome and correction models, but not the treatment anchor.

4.3 A FITTED-ANCHOR VALIDITY GUARANTEE

Evaluate Equation (12) with fitted functions and $\widehat{A}_0$ to obtain $\widehat{\psi}_j$ on n test observations. Let $\overline{\psi}_j$ and s_j be its sample mean and standard deviation. Fix the number of transformations J and a positive standard-deviation floor τ. Write Z_j for the standardized magnitude of the estimated moment and $p_{Y,\mathrm{corr}}$ for the corrected outcome p-value. The joint p-value $p_{\mathrm{AO,corr}}$ combines it with the treatment-relevance p-value p_T. Using the standard normal distribution function Φ, define

$$Z_j = \frac{\sqrt{n}|\overline{\psi}_j|}{\max(s_j, \tau)}, \quad p_{Y,\mathrm{corr}} = \min\{1, 2J[1 - \Phi(\max_j Z_j)]\}, \quad p_{\mathrm{AO,corr}} = \max(p_T, p_{Y,\mathrm{corr}}). \tag{13}$$

The joint value requires both comparisons to reject. The factor J handles multiple transformations; τ prevents division by zero.

For $p = 2$ or 4, write $\|f\|_p = \{\mathbb{E}[|f(O)|^p]\}^{1/p}$ and $f \in L_p$ when this norm is finite. For fitted functions the expectation uses a fresh record conditional on development data. A regression error is the fitted function minus its population counterpart. For regression errors, these norms give root-mean-square error and the fourth root of mean fourth-power error, respectively. The notation $e_n = o_{\mathbb{P}}(a_n)$ means that e_n/a_n tends to zero in probability. The $o_{\mathbb{P}}(n^{-1/4})$ rates make products of two errors negligible compared with $n^{-1/2}$ sampling fluctuations.

Theorem 1 (Validity with an estimated anchor)**.** *Suppose the test observations are i.i.d. and independent of development data, J is fixed, and each h_j has bounded first and second derivatives. Assume $R_Y, \psi_j \in L_2$, $\|\widehat{\psi}_j - \psi_j\|_2 = o_{\mathbb{P}}(1)$, and each regression error in $m_1, m_0, g, r_j, q_j, k_j$ is $o_{\mathbb{P}}(n^{-1/4})$ in L_4. Under $\theta_{Y,j}(A_0) = 0$ for all j, for every fixed $\alpha \in (0, 1)$,*

$$\limsup_{n\to\infty} \mathbb{P}(p_{\mathrm{AO,corr}} \le \alpha) \le \alpha. \tag{14}$$

Under Assumption 2, this includes the restricted no-hidden-confounding model.

At level .05, the false-rejection probability has asymptotic upper bound 5% when all ideal outcome moments vanish; finite samples need not meet that bound. Appendix A.5 gives weaker product-rate conditions and verifies the Gaussian ordinary-least-squares case with $h_1(a) = a$. Validity holds for each fixed law, including zero-variance scores; full proxy inputs alone do not ensure the learning rates.

Input: prespecified proxies, contexts, transformations, and independent development/test samples.

1. Split development data into two balanced folds. Fit any representation and m_1, m_0, g on one fold; predict the other. Swap folds.
2. Use the pooled out-of-fold predictions to train $\widehat{r}_j, \widehat{q}_j, \widehat{k}_j$ and treatment residual regressions. Refit the representation and m_1, m_0, g on all development data.
3. On test data, form $\widehat{A}_0 = \widehat{m}_1 - \widehat{m}_0$ and the corrected products in Equation (12); compute p_T and $p_{Y,\mathrm{corr}}$.
4. Report both components and their maximum. For independently selected graph queries, apply Holm's correction across candidates.

Algorithm 1: AO-CES. Test observations are never used for fitting or tuning.

5 EXPERIMENTS

We evaluate the validity and power of AO-CES. *Plug-in AO-CES* uses the estimated anchor directly; *corrected AO-CES* adds the adjustment in Section 4. Both versions share fitted predictions and test observations. Unless noted, each condition uses 100 independent datasets, an 80/20 development/test split of rows, and significance level .05.

Data and compared predictors. Continuous simulations combine latent factors and independent errors to generate T, Y and noisy proxies; Y also depends on T. Null conditions remove shared latent contributions but retain the treatment effect. Mechanisms include Gaussian and Laplace linear models, nonlinear models, separate factors, sine/cosine proxies, and rotated factors (Appendix C). Synthetic

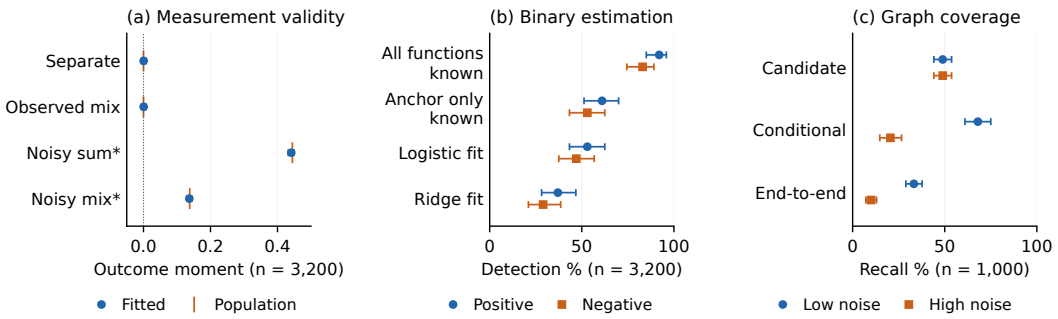

Figure 2: What limits corrected AO-CES detection? (a) Without a shared cause, starred measurement schemes nevertheless yield a nonzero outcome moment: the measurement assumptions fail. (b) Binary detection using ridge or logistic fits, an exact anchor, or all population functions known, on matched test observations. (c) Candidate recall counts confounded pairs selected by FCI; conditional recall counts detections among those selected; overall recall includes missed pairs. Error bars: 95% Monte Carlo mean intervals (a), Wilson intervals (b), and whole-graph bootstrap intervals (c).

samples have $n = 800$. Wisconsin Diagnostic Breast Cancer (WDBC) supplies 569 real covariate records for synthetic T, Y; diagnosis labels are unused (Street et al., 1993). Five binary conditions draw T, Y and 20 proxies from probabilities determined by a four-category latent variable (Lee, 2014). Seven mechanisms in both confounding states, two rotated-factor states, and five binary populations give 21 conditions (Appendix G).

The six AO-CES treatment predictors use full $\boldsymbol{W}$ alone, or append two features from principal component analysis (PCA), treatment-predictive partial least squares (PLS), a reconstruction variational autoencoder (VAE), a treatment-supervised VAE, or a treatment variational information bottleneck (VIB) (Jolliffe, 2002; Wold et al., 2001; Kingma & Welling, 2014; Alemi et al., 2017). All retain the original measurements: these are AO-CES feature-learning ablations. The 12,600 ridge-regression fits report identity and smooth tests separately. Gaussian calibration studies use ordinary least squares.

Measurement assumptions and fitting error. Generate $T = U_T + \varepsilon_T$ and $Y = T + .8HU_T + U_Y + \varepsilon_Y$, with independent Gaussian roots and $H = 0/1$ switching confounding off/on (Equation (I.1)). In a 1,600-fit comparison, separate noisy measurements and their invertible transformation give identical corrected AO-CES results. A noisy sum and a latent mixture with newly added sensor noise instead violate the ideal outcome null even at $H = 0$. Their observed moments match the nonzero predictions (Figure 2a). Estimation correction cannot repair these invalid measurement assumptions.

A separate Gaussian study adds shrinking outcome-only anchor error. At $n = 3200$, corrected AO-CES reduces plug-in false positives from 90% to 5.5% while retaining 100% detection (Appendix F).

Detection and learning. Correction changes full-proxy identity-test detection from 48% to 65% for rotated factors and from 96% to 79% for WDBC. Added features give no uniform gain; at $n = 800$, binary detection is only 0–3%. To explain this, 1,000 matched binary datasets compare fitted AO-CES, an exact anchor with fitted remaining regressions, and all population functions known. Oracles use observed proxies, not realized latent values. At $n = 3200$, knowing all population functions raises detection from 29–37% to 83–92%; even known functions reach only 27–32% at $n = 800$. A prespecified logistic follow-up reaches 47–53% at $n = 3200$ but only 2% at $n = 800$. Small test samples and regression errors limit power (Figure 2b; Appendix J).

Published diagnostics. Lee's binary proxy test detects 90–92/99 completed alternatives but rejects 14/98 outcome-only and 20/98 treatment-only cases (Lee, 2014). It uses all 800 rows, whereas AO-CES reserves 160 for testing. Higher-order least squares (HOLS) and the kernel-regression diagnostic KRCD test regression-model restrictions (Schultheiss et al., 2024; Chen et al., 2026); Liu's proxy test concerns a remaining causal relation (Liu et al., 2024), which our no-confounding simulations retain. Appendices D.1–E.1 separate these results from AO-CES ablations; effect estimators such as CEVAE and TEDVAE are not converted into confounding tests (Louizos et al., 2017; Zhang et al., 2021).

5.1 APPLICATION: ADDING PROXY EVIDENCE TO FCI

Fast Causal Inference (FCI) returns a partial ancestral graph (PAG) (Colombo et al., 2012; Zhang, 2008). Its marks describe ancestry constraints: an arrowhead at Y excludes Y being an ancestor of the other endpoint; a circle leaves that endpoint unresolved. A *bow* instead represents a direct-effect arrow and a shared-cause edge together. AO-CES adds a common-cause annotation without changing PAG endpoints or establishing a direct effect.

Data and workflow. Each observation records ten primary variables: four treatments, four outcomes, and two baseline covariates. There are $4 \times 4 = 16$ treatment–outcome queries and 16 additional proxies. The 600 structured datasets contain four direct treatment–outcome pairs, with zero, two, or four sharing a hidden cause, at $n = 1000/3000$: 100 replicates per configuration (Equation (H.1)). The randomized study draws direct and hidden links independently, allowing either, both, or neither (Equation (K.1)). Its 100 null and 100 mixed graphs, each with proxy-noise standard deviations $.5$ and 1.5, give 400 datasets with $n = 1000$.

FCI (Zheng et al., 2024) sees *25% of observation rows*, keeping all ten primary variables; it sees no proxies. No directions are imposed on FCI. Its treatment–outcome adjacencies select queries for AO-CES, regardless of endpoint marks. The remaining rows provide 60% of the original sample for fitting and 15% for testing with full-proxy least squares. Baseline contexts and treatment/outcome roles are prespecified. Holm correction accounts for multiple tests (Holm, 1979); independent selection preserves asymptotic annotation-error control (Lemma H.1).

What is scored. Precision is the fraction of annotations on truly confounded pairs; recall is the fraction of all truly confounded queries annotated. Conditional recall uses only confounded pairs selected by FCI. Familywise error is the probability of at least one false annotation in a dataset. The unchanged PAG has unchanged structural Hamming distance (SHD), which cannot measure added annotations. Appendix H separately reports adjacency F1.

Results. In structured graphs, corrected AO-CES reduces all-null familywise error from plug-in 13%/11% to 4%/3% at $n = 1000/3000$, with 93–100% recall. Randomization exposes missed candidates: FCI selects 200 of 409 confounded pairs (48.9%) per sensor-noise condition. With lower-noise proxies, corrected AO-CES has 99.3% precision, 68.0% conditional recall, and 33.3% overall recall; higher noise lowers overall recall to 10.0%. All-null familywise error is 3%/1% (Figure 2c). Of 136 lower-noise true annotations, 69 involve a circle endpoint, illustrating added common-cause evidence without orientation.

Testing all 16 queries avoids missed candidates but increases the multiplicity penalty. Its low/high-noise recall is 37.2%/6.6%, versus FCI selection's 33.3%/10.0%: neither strategy uniformly wins (Appendices K–L).

6 DISCUSSION AND LIMITATIONS

Reliable proxy evidence requires both an appropriate measurement model and a valid fitted test. Retaining full proxy inputs prevents information loss from compression, but cannot undo a misleading mixture already present in the measurements. The correction removes first-order fitting bias; it does not change that population target or guarantee accurate learning for every predictor. The binary diagnosis makes the distinction concrete: better treatment prediction helps at larger samples, while noisy correction regressions and weak test-sample signal still limit detection.

The graph application adds evidence about hidden common causes to relationships selected by FCI. Screening reduces the multiplicity penalty but can exclude detectable relations; its benefit depends on proxy quality and candidate coverage, not just the accuracy of the local test. In the treatment-access example, a positive result motivates better measurement of household resources, an adjustment review, or sensitivity analysis—not automatic adjustment for the anchor. A negative result does not exclude weakly measured or cancelling confounding. The contribution is a diagnostic whose measurement assumptions and fitting error can be examined separately, not a certificate that an observational effect is identified.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, 1996. URL https://www.public.asu.edu/~rlberge1/papers/statsci96.pdf.

Yikai Chen, Yunxin Mao, Chunyuan Zheng, Hao Zou, Shanzhi Gu, Shixuan Liu, Yang Shi, Wenjing Yang, Kun Kuang, and Haotian Wang. Detecting unobserved confounders: A kernelized regression approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20381–20389, 2026. doi: 10.1609/aaai.v40i24.39127.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6): 2786–2819, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022. doi: 10.3982/ECTA16294.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

Piersilvio De Bartolomeis, Javier Abad Martinez, Konstantin Donhauser, and Fanny Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1045–1053, 2024. URL https://proceedings.mlr.press/v238/de-bartolomeis24a.html.

Shachi Deshpande, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Felix Elwert and Fabian T. Pfeffer. The future strikes back: Using future treatments to detect and reduce hidden bias. *Sociological Methods & Research*, 51(3):1014–1051, 2022. doi: 10.1177/0049124119875958.

Juan Carlos Escanciano and Telmo Pérez-Izquierdo. Automatic locally robust GMM with machine-learning-generated regressors. arXiv preprint arXiv:2301.10643v4; first posted 2023, 2026. URL https://arxiv.org/abs/2301.10643v4.

Zijian Guo, Jing Cheng, Scott A. Lorch, and Dylan S. Small. Using an instrumental variable to test for unmeasured confounding. *Statistics in Medicine*, 33(20):3528–3546, 2014. doi: 10.1002/sim.6227.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):20170013, 2018. doi: 10.1515/jci-2017-0013.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002. doi: 10.1007/b98835.

Hamid Kalantari, Russell Greiner, and Pouria Ramazi. Extendable and iterative structure learning strategy for bayesian networks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3n6DYH3cIP.

Rickard Karlsson and Jesse H. Krijthe. Falsification of unconfoundedness by testing independence of causal mechanisms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29128–29147, 2025. URL https://proceedings.mlr.press/v267/karlsson25a.html.

Rickard K. A. Karlsson and Jesse H. Krijthe. Detecting hidden confounding in observational data using multiple environments. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6114.

Haley Colgate Kottler and Amy Cochran. A proxy-based approach for unmeasured confounding in electronic health records research. arXiv preprint arXiv:2506.12177v2; first posted 2025, 2026. URL https://arxiv.org/abs/2506.12177v2.

Erich Kummerfeld, Jaewon Lim, and Xu Shi. Data-driven automated negative control estimation (DANCE): Search for, validation of, and causal inference with negative controls. *Journal of Machine Learning Research*, 25(229):1–35, 2024. URL https://jmlr.org/papers/v25/22-1062.html.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1986.

Wen-Chung Lee. Detecting and correcting the bias of unmeasured factors using perturbation analysis: a data-mining approach. *BMC Medical Research Methodology*, 14:18, 2014. doi: 10.1186/1471-2288-14-18.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Furui Liu and Laiwan Chan. Confounder detection in high-dimensional linear models using first moments of spectral measures. *Neural Computation*, 30(8):2284–2318, 2018. doi: 10.1162/neco_a_01099.

Mingzhou Liu, Xinwei Sun, Yu Qiao, and Yizhou Wang. Causal discovery via conditional independence testing with proxy variables. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31866–31889, 2024.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Anton Rask Lundborg, Ilmun Kim, Rajen D. Shah, and Richard J. Samworth. The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52 (6):2851–2878, 2024. doi: 10.1214/24-AOS2447. URL https://arxiv.org/abs/2211.02039.

Takashi Nicholas Maeda and Shohei Shimizu. RCD: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 735–745, 2020.

Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 97–106, 2021.

Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. Identifying confounding from causal mechanism shifts. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4897–4905, 2024.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7512–7523, 2021.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038.

Chan Park, David B. Richardson, and Eric J. Tchetgen Tchetgen. Single proxy control. *Biometrics*, 80(2):ujae027, 2024. doi: 10.1093/biomtc/ujae027.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, 2017.

Drago Plečko, Patrik Okanović, Torsten Hoefler, and Elias Bareinboim. Confounder detection via treatment intent: A new observational study design. arXiv preprint arXiv:2605.26413, 2026. URL https://arxiv.org/abs/2605.26413.

Abbavaram Gowtham Reddy and Vineeth N. Balasubramanian. Detecting and measuring confounding using causal mechanism shifts. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/711ebaa5fcc8321d0075c1b42870b425-Abstract-Conference.html.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/21c5bba1dd6aed9ab48c2b34c1a0adde-Abstract.html.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Vahideh Sanjaroonpouri and Pouria Ramazi. Linear SCM identification in the presence of confounders and gaussian noise. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjxuqI4KwU.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022. URL https://jmlr.org/papers/v23/21-1328.html.

Christoph Schultheiss, Peter Bühlmann, and Ming Yuan. Higher-order least squares: Assessing partial goodness of fit of linear causal models. *Journal of the American Statistical Association*, 119(546):1019–1031, 2024. doi: 10.1080/01621459.2022.2157728.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

W. Nick Street, William H. Wolberg, and Olvi L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pp. 861–870, 1993.

Victor Veitch, Yixin Wang, and David M. Blei. Using embeddings to correct for unobserved confounding in networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001. doi: 10.1016/S0169-7439(01)00155-1.

Yong Wu, Yanwei Fu, Shouyan Wang, and Xinwei Sun. Bivariate causal discovery with proxy variables: Integral solving and beyond. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67202–67259, 2025.

Feng Xie, Zhengming Chen, Shanshan Luo, Wang Miao, Ruichu Cai, and Zhi Geng. Automating the selection of proxy variables of unmeasured confounders. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54430–54459, 2024.

Myeonghun Yu, Xu Shi, and Eric J. Tchetgen Tchetgen. Fortified proximal causal inference with many invalid proxies. arXiv preprint arXiv:2506.13152, 2025. URL https://arxiv.org/abs/2506.13152.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9(47): 1437–1474, 2008. URL https://www.jmlr.org/papers/v9/zhang08a.html.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10923–10930, 2021. doi: 10.1609/aaai.v35i12.17304.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in Python. *Journal of Machine Learning Research*, 25(60):1–8, 2024. URL https://www.jmlr.org/papers/v25/23-0970.html.

AI USE STATEMENT

Generative AI tools assisted with manuscript restructuring, language editing, literature organization, proof checking, experiment-code implementation and review, experiment orchestration, figure refinement, and typesetting. The authors are responsible for the claims, citations, code, results, and final content.

ETHICS STATEMENT

We collected no new human-subject data and used only public benchmark datasets; the diagnostic should not replace randomized evidence, domain review, or a formal causal-effect identification analysis.

A NOTATION AND SUPPORTING ARGUMENTS

Table 1: Notation and the medication example. All symbols are introduced where used in the main text.

Symbol	Meaning
T, Y	Treatment and outcome; receiving medication ($T = 1$ versus 0) and a 0–100 recovery score.
$\boldsymbol{C}_T$	Recorded context used to predict T, e.g. age and site.
$\boldsymbol{C}_Y$	Recorded context used with T to predict Y, e.g. age, site, and comorbidities. $\boldsymbol{C}$ collects both sets without duplicates.
$\boldsymbol{W} \in \mathbb{R}^d$	The d proxy measurements, e.g. a noisy income record ($d = 1$). These are available to the analyst, not necessarily used to assign treatment.
$\boldsymbol{U}$	Relevant pretreatment factors not directly measured. A factor influencing T and Y through a route bypassing T is a hidden common cause.
$\boldsymbol{Z} = r(\boldsymbol{W})$	Learned proxy features, optionally added to $\boldsymbol{W}$ when predicting T. Compressed ablations instead use $\boldsymbol{Z}$ alone.
$A_0, \widehat{A}_0$	The ideal treatment anchor and its estimate: scalar changes in treatment prediction.

A.1 OBSERVATIONAL NON-IDENTIFIABILITY

Observational data alone need not distinguish a confounded effect from an unconfounded one (Pearl, 2009; Peters et al., 2017; Sanjaroonpouri & Ramazi, 2025). The following standard Gaussian construction illustrates the issue, not a new identification result.

Let U and all disturbances below be mutually independent standard-Gaussian variables. The unconfounded model

$$T = \varepsilon_T, \qquad Y = .5T + \sqrt{.75}\,\varepsilon_Y$$

has unit marginal variances and covariance .5. The confounded model

$$T = .5U + \sqrt{.75}\,\widetilde{\varepsilon}_T, \qquad Y = .35T + .30U + \sqrt{.6825}\,\widetilde{\varepsilon}_Y$$

also has unit marginal variances and covariance .5. The observed pairs therefore have the same bivariate Gaussian law. Yet setting $T = t$ gives mean outcomes $.5t$ and $.35t$, respectively. The second model has the shared cause U, and its observational regression coefficient .5 differs from its direct structural coefficient .35.

A.2 PROJECTION IDENTITIES AND SEPARATED MEASUREMENTS

These are standard conditional-expectation calculations underlying the projection approach (Lundborg et al., 2024). With the definitions of m_1, m_0 in Section 3.2, the tower property gives

$$\mathbb{E}[A_0 \mid \boldsymbol{C}_T] = \mathbb{E}[\mathbb{E}[T \mid \boldsymbol{W}, \boldsymbol{C}_T] \mid \boldsymbol{C}_T] - \mathbb{E}[T \mid \boldsymbol{C}_T] = 0.$$

If $\boldsymbol{C}_T$ is a measurable function of $\boldsymbol{C}_Y$, m_0 is fixed when conditioning on $(T, \boldsymbol{C}_Y)$, hence

$$A_0 - \mathbb{E}[A_0 \mid T, \boldsymbol{C}_Y] = m_1 - \mathbb{E}[m_1 \mid T, \boldsymbol{C}_Y].$$

Thus context centering cancels from the linear outcome residual, not from every nonlinear transformation of the anchor.

For square-integrable T, Y, sufficient separation conditions are

$$\mathbb{E}[T \mid \boldsymbol{W}_T, \boldsymbol{W}_Y, \boldsymbol{C}_T] = \mathbb{E}[T \mid \boldsymbol{W}_T, \boldsymbol{C}_T], \qquad \mathbb{E}[Y \mid T, \boldsymbol{C}_Y, \boldsymbol{W}_T] = \mathbb{E}[Y \mid T, \boldsymbol{C}_Y]. \tag{A.1}$$

If $\boldsymbol{C}_T$ is measurable from $\boldsymbol{C}_Y$, the first equality makes A_0 a function of $(\boldsymbol{W}_T, \boldsymbol{C}_T)$, and the second then yields $\mathbb{E}[Y \mid T, \boldsymbol{C}_Y, A_0] = \mathbb{E}[Y \mid T, \boldsymbol{C}_Y]$ by iterated expectation. Multiplying this zero conditional residual mean by any square-integrable $h(A_0)$ proves

$$\mathbb{E}[\{Y - \mathbb{E}[Y \mid T, \boldsymbol{C}_Y]\}\{h(A_0) - \mathbb{E}[h(A_0) \mid T, \boldsymbol{C}_Y]\}] = 0. \tag{A.2}$$

For a fixed compression $\boldsymbol{Z} = r(\boldsymbol{W})$, put $A_Z = \mathbb{E}[T \mid \boldsymbol{Z}, \boldsymbol{C}_T] - \mathbb{E}[T \mid \boldsymbol{C}_T]$. The same conclusion holds if

$$\mathbb{E}[T \mid \boldsymbol{Z}, \boldsymbol{C}_T] = \mathbb{E}[T \mid \boldsymbol{W}_T, \boldsymbol{C}_T] \quad \text{almost surely.} \tag{A.3}$$

because then $A_Z = A_0$; this is a condition, not a guarantee for a learned representation. The tower property and the orthogonality of conditional-expectation residuals also give

$$A_Z = \mathbb{E}[A_0 \mid \boldsymbol{Z}, \boldsymbol{C}_T], \qquad \mathbb{E}[A_0^2] = \mathbb{E}[A_Z^2] + \mathbb{E}[(A_0 - A_Z)^2]. \tag{A.4}$$

The last term is treatment information lost by compression; it follows by expanding $A_0 = A_Z + (A_0 - A_Z)$ and using $\mathbb{E}[A_Z(A_0 - A_Z)] = 0$.

A.3 The sharp compression bound

Proposition A.1 (A sharp compression bound)**.** *For the linear, unclipped moment, let T, Y be square-integrable, $\boldsymbol{Z} = r(\boldsymbol{W})$, and $\theta_{Y,1}(A_0) = 0$. Write $v = \mathbb{E}[A_0^2] > 0$, $s = \mathbb{E}[A_Z^2]$, and $\sigma_Y^2 = \mathbb{E}[R_Y^2]$. Then*

$$|\theta_{Y,1}(A_Z)| \le \sigma_Y \sqrt{\frac{s(v-s)}{v}}. \tag{A.5}$$

The bound is attained for every retained fraction $s/v \in (0,1)$ by a model with separate treatment-only and outcome-only factors and no hidden common cause.

Here s is retained treatment signal and $v - s$ is its loss. Full retention preserves the outcome null, but 99% retention still permits a moment of about .10 when $v = \sigma_Y^2 = 1$.

The projection identity gives $\mathbb{E}[A_0 A_Z] = s$. Since $\mathbb{E}[R_Y \mid T, \boldsymbol{C}_Y] = 0$, residualizing the score does not change its product expectation with R_Y. The assumed ideal null therefore gives

$$\theta_{Y,1}(A_Z) = \mathbb{E}\Big[R_Y \left(A_Z - \frac{s}{v} A_0\right)\Big], \qquad \mathbb{E}\left[\left(A_Z - \frac{s}{v} A_0\right)^2\right] = s - \frac{s^2}{v}.$$

Cauchy–Schwarz proves Equation (A.5). To show sharpness without hidden confounding, take constant contexts, let U_T, U_Y, ε_T be independent standard normals, and set $T = U_T + \varepsilon_T$, $Y = \beta T + U_Y$, and $\boldsymbol{W} = (U_T, U_Y)$. Choose $\beta \neq 0$ to retain the direct treatment effect. For any $q \in (0,1)$, retain the single coordinate $Z = \sqrt{q}\, U_T + \sqrt{1-q}\, U_Y$. Gaussian conditioning gives

$$A_0 = U_T, \quad A_Z = qU_T + \sqrt{q(1-q)}\, U_Y, \quad v = \sigma_Y^2 = 1, \quad s = q, \quad \theta_{Y,1}(A_Z) = \sqrt{q(1-q)}.$$

Thus equality holds throughout the retained-signal range, although U_T and U_Y are not a shared cause. Adding $Z_\perp = -\sqrt{1-q}\, U_T + \sqrt{q}\, U_Y$ makes the representation invertible and recovers A_0 exactly. Discarding one mixed coordinate, rather than the rotation itself, causes the population failure. No finite-sample rotation invariance is asserted for the fitted regressors. □

A.4 Treatment prediction error and outcome contamination

Let $U_T, U_Y, \varepsilon_T, \varepsilon_Y$ be independent standard normal variables and set

$$\boldsymbol{W} = (U_T, U_Y), \qquad T = U_T + \varepsilon_T, \qquad Y = \beta T + U_Y + \varepsilon_Y.$$

There is no hidden common cause of T and Y, and the ideal anchor without context is $A_0 = U_T$. For an approximation $A_\delta = U_T + \delta U_Y$, where δ is a scalar fitting error,

$$\mathbb{E}[(T - A_\delta)^2] - \mathbb{E}[(T - A_0)^2] = \delta^2.$$

Yet $\mathbb{E}[Y \mid T] = \beta T$ and $\mathbb{E}[A_\delta \mid T] = T/2$, so the linear outcome moment is

$$\begin{aligned}\theta_{Y,1}(A_\delta) &= \mathbb{E}[(U_Y + \varepsilon_Y)(U_T - T/2 + \delta U_Y)] \\ &= \delta, \qquad \theta_{Y,1}(A_0) = 0.\end{aligned}$$

Thus prediction error enters quadratically while outcome contamination enters linearly. For $\delta = n_{\text{test}}^{-1/4}$, prediction error vanishes but the root-n_{test} outcome-moment shift diverges. For the uncorrected, clipped procedure, one sufficient bridge to the ideal null would be

$$\max_{j \le J} |\theta_{Y,j}(\widehat{A}_{\text{plug}}) - \theta_{Y,j}(A_0^\kappa)| = o_{\mathbb{P}}(n_{\text{test}}^{-1/2}), \qquad A_0^\kappa = \text{clip}(A_0, -\kappa, \kappa). \tag{A.6}$$

Here $\widehat{A}_{\text{plug}}$ is the clipped plug-in anchor of Equation (B.2); predictive accuracy alone does not establish this condition. The corrected statistic instead cancels the first-order shift, as shown next.

A.5 Proof of fitted-anchor validity

Use the regressions and corrected product from Equations (11)–(12). Conditional centering of $T - m_1$ and $T - m_0$ gives $\mathbb{E}[\psi_j] = \theta_{Y,j}(A_0)$. Here h_j denotes the *entire* smooth transformation, including any smooth saturation: the corrected implementation uses $s(a) = 4\tanh(a/4)$ before its transformation bank, with the derivative computed by the chain rule. The identity transformation without saturation is also covered. The hard clipping used in the earlier experiments is not covered by the smoothness argument below; it would require additional conditions at its two boundaries.

Product-rate conditions. Retain the independence, fixed-J, smoothness, moment, and score-consistency assumptions of Theorem 1, but replace its individual regression-rate requirements by the following weaker condition. Let $d = \widehat{m}_1 - m_1$, $e = \widehat{m}_0 - m_0$, $\Delta = d - e$, $u = \widehat{g} - g$, $v_j = \widehat{r}_j - r_j$, $\xi_j = \widehat{q}_j - q_j$, and $\zeta_j = \widehat{k}_j - k_j$. Suppose the following bound is $o_{\mathbb{P}}(n^{-1/2})$ for every j:

$$\frac{\|h_j''\|_\infty}{2}\|R_Y\|_2\|\Delta\|_4^2 + \|h_j'\|_\infty\|u\|_2\|\Delta\|_2 + \|u\|_2\|v_j\|_2 + \|\xi_j\|_2\|d\|_2 + \|\zeta_j\|_2\|e\|_2. \tag{A.7}$$

For $\widehat{\theta}_j = n^{-1}\sum_i \widehat{\psi}_j(O_i)$,

$$\sqrt{n}\{\widehat{\theta}_j - \theta_{Y,j}(A_0)\} = n^{-1/2}\sum_i \{\psi_j(O_i) - \theta_{Y,j}(A_0)\} + o_{\mathbb{P}}(1). \tag{A.8}$$

Here $\widehat{\theta}_j = \overline{\psi}_j$ in Equation (13). Under the ideal outcome null, that calibration is asymptotically valid even with zero score variances. Together with Assumption 2, this gives the claimed causal-null interpretation.

Proof. Suppress j and condition on development data. Conditional expectations give $\mathbb{E}[R_Y \mid V] = 0$, $\mathbb{E}[T - m_1 \mid X] = 0$, and $\mathbb{E}[T - m_0 \mid C] = 0$. Expanding the fitted score therefore gives the exact remainder

$$\begin{aligned}\mathbb{E}[\widehat{\psi}] - \theta_Y(A_0) = \mathbb{E}[R_Y\{h(A_0 + \Delta) - h(A_0) - h'(A_0)\Delta\}] \\ - \mathbb{E}[u\{h(A_0 + \Delta) - h(A_0)\}] + \mathbb{E}[uv] - \mathbb{E}[\xi d] + \mathbb{E}[\zeta e].\end{aligned} \tag{A.9}$$

In particular, $\mathbb{E}[R_Y h'(A_0)\Delta] = \mathbb{E}[qd] - \mathbb{E}[ke]$ cancels the corresponding linear errors from the added terms. Taylor's theorem and Cauchy–Schwarz bound the remainder by Equation (A.7); every first-order nuisance derivative vanishes. Conditional L_2 score consistency makes the centered empirical difference between fitted and population scores $o_{\mathbb{P}}(n^{-1/2})$, proving Equation (A.8). The finite-variance central limit theorem and variance consistency apply to the population scores. For positive variance, replacing its standard deviation by its maximum with τ can only shrink the limiting normal statistic. For zero variance, Equation (A.8) and the fixed positive floor give $Z_j = o_{\mathbb{P}}(1)$ under the null. Bonferroni's inequality proves the stated calibration without assumptions about dependence across transformations. Finally, $\{\max(p_T, p_{Y,\text{corr}}) \le \alpha\} \subseteq \{p_{Y,\text{corr}} \le \alpha\}$. □

The implementation sets $\tau = .01$ in the units of each reported generator. The floor is fixed as sample size grows; it trades power for protection against degenerate scores. The L_4 rates in Theorem 1 imply Equation (A.7) by Hölder's inequality, completing its proof; these rates must be established for the chosen regressions.

Corollary A.1 (A finite-dimensional case)**.** *Suppose the observed coordinates are jointly Gaussian, dimensions are fixed, $h(a) = a$, and development and test sample sizes have a positive limiting ratio. Assume nonsingular regressor Gram matrices after redundant or constant columns are removed; constant contexts are allowed. Ordinary least-squares regressions for m_1, m_0, g, r, q, k, with a fixed number of balanced inner folds for generated targets, satisfy Theorem 1's conditions. Thus the corrected test is asymptotically valid under its ideal outcome null.*

All required conditional means are affine in this case. First-stage coefficient errors are $O_{\mathbb{P}}(n^{-1/2})$. Here $O_{\mathbb{P}}(n^{-1/2})$ means that multiplying the error by $\sqrt{n}$ gives a quantity bounded in probability. Each inner-fold target perturbation is a fixed-dimensional regressor multiplied by such a coefficient error; averaging over a fixed number of folds preserves that rate in the downstream least-squares fits. Gaussian moments then give L_2 score consistency and an $O_{\mathbb{P}}(n^{-1})$ remainder bound. When $\boldsymbol{C}_T$ is contained in $\boldsymbol{C}_Y$, $k = 0$ for the identity transformation and m_0 cancels after outcome residualization.

The cancellation is visible in Appendix A.4's example. There $q = U_Y$, $m_0 = 0$, and an imperfect predictor $\widehat{m}_1 = U_T + \delta U_Y$ gives an uncorrected outcome mean δ. The added mean is $\mathbb{E}[U_Y\{T - \widehat{m}_1\}] = -\delta$, cancelling it exactly. If instead $\widehat{q} = (1+\epsilon)U_Y$, the remaining bias is $-\epsilon\delta$, with the other nuisances correct for the fitted score. The correction therefore changes a first-order contamination error into a product error.

A compressed predictor need not converge to $m_1(X)$; supplying full $\boldsymbol{W}$ avoids that information loss but does not prove the learning rates or causal assumptions.

A.6 Measurement geometry and primitive sufficient conditions

Proof of Proposition 1. Write $\Omega = \mathrm{Var}(\boldsymbol{\varepsilon}_W)$, $\sigma_T^2 = \mathrm{Var}(\varepsilon_T)$, and $D = \|\boldsymbol{a}\|^2 + \sigma_T^2$. Gaussian conditioning gives $\boldsymbol{M} = \mathbb{E}[\boldsymbol{U} \mid \boldsymbol{W}] = L^\top(LL^\top + \Omega)^{-1}\boldsymbol{W}$, $\mathrm{Var}(\boldsymbol{M}) = Q = L^\top(LL^\top + \Omega)^{-1}L$, and $A_0 = \boldsymbol{a}^\top \boldsymbol{M}$. The centered outcome residual is

$$R_Y = \boldsymbol{b}^\top \boldsymbol{U} - \frac{\boldsymbol{a}^\top \boldsymbol{b}}{D} T + \varepsilon_Y.$$

Because $\mathbb{E}[R_Y \mid T] = 0$, the linear moment equals $\mathbb{E}[R_Y A_0]$. The identities $\mathrm{Cov}(\boldsymbol{U}, \boldsymbol{M}) = Q$ and $\mathrm{Cov}(T, A_0) = \boldsymbol{a}^\top Q \boldsymbol{a}$ establish Equation (10).

The residuals R_Y and $B = A_0 - \mathbb{E}[A_0 \mid T]$ are jointly Gaussian and jointly independent of T. If $\mathrm{Var}(B) > 0$, their covariance is zero exactly when the conditional mean of R_Y given (T, B) is zero. If $\mathrm{Var}(B) = 0$, A_0 is a function of T and both conclusions hold automatically. This proves the conditional-mean equivalence, including degenerate anchors.

Holding L, Ω fixed, take disjoint latent index sets I_T, I_Y, restrict $\boldsymbol{a}$ to I_T and $\boldsymbol{b}$ to I_Y. Then $\boldsymbol{a}^\top \boldsymbol{b} = 0$, so the moment reduces to $\boldsymbol{b}^\top Q \boldsymbol{a}$. It is zero for every supported coefficient pair if $Q_{I_Y, I_T} = 0$. Conversely, take $\boldsymbol{a}$ and $\boldsymbol{b}$ to be coordinate vectors in the respective blocks; every cross-block entry must then vanish. The necessity statement concerns validity for all coefficient pairs, not the absence of accidental cancellation for one pair. □

Primitive measurement conditions. Suppose the latent roots are partitioned into independent blocks and each measurement channel depends linearly on only its corresponding block. If the measurement errors are also independent between these blocks, L and Ω are block diagonal after reordering. Their covariance formula makes Q block diagonal as well, proving the required null whenever the treatment and outcome coefficients occupy different blocks. More generally, outside the Gaussian family, take latent blocks U_T, U_Y and exogenous errors all mutually independent, let $T = f(U_T, \varepsilon_T)$, $Y = g(T, U_Y, \varepsilon_Y)$, and observe separate channels $W_T = s_T(U_T, \eta_T)$, $W_Y = s_Y(U_Y, \eta_Y)$. Assuming integrability, the independence of the two blocks yields $\mathbb{E}[T \mid W_T, W_Y] = \mathbb{E}[T \mid W_T]$ and $\mathbb{E}[Y \mid T, W_T] = \mathbb{E}[Y \mid T]$. These imply the ideal-anchor null by iterated expectation. This is a structural sufficient class for the separation conditions in Equation (A.1); arbitrary mixing measurements are not included.

If $\widetilde{\boldsymbol{W}} = R\boldsymbol{W}$ for a nonsingular square R, then $\sigma(\widetilde{\boldsymbol{W}}) = \sigma(\boldsymbol{W})$ and the anchor and Q are unchanged. In the Gaussian formula, both loadings and measurement covariance transform: $\widetilde{L} = RL$, $\widetilde{\Omega} = R\Omega R^\top$. This differs from mixing latent signals and then adding independent sensor noise. For

example,

$$L = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \Omega = \tfrac{1}{4} I_2, \qquad Q = \frac{1}{29} \begin{pmatrix} 20 & 4 \\ 4 & 24 \end{pmatrix}.$$

For $\boldsymbol{a} = (1, 0)^\top$ and $\boldsymbol{b} = (0, 1)^\top$, this full-rank measurement model has $\theta_{Y,1} = 4/29 > 0$ without a hidden common cause. All examples satisfy proxy mean exclusion because the independent outcome error supplies the only residual outcome randomness conditional on $(T, \boldsymbol{U})$.

Scalar measurement-quality specialization. For a scalar $U \sim N(0, s)$, independent Gaussian errors, and $T = aU + \varepsilon_T$, $Y = \beta T + bU + \varepsilon_Y$, $W = \lambda U + \varepsilon_W$, write $q = \mathrm{Var}\{\mathbb{E}[U \mid W]\} = \lambda^2 s^2/(\lambda^2 s + \omega)$, where the measurement denominator is positive, $\mathrm{Var}(T) > 0$, and $\omega = \mathrm{Var}(\varepsilon_W)$. The same covariance calculation gives

$$\theta_{Y,1}(A_0) = abq \frac{\sigma_T^2}{a^2 s + \sigma_T^2}. \tag{A.10}$$

For $ab \neq 0$, positive q and positive treatment-noise variance produce a nonzero moment. Weak measurement information or vanishing residual treatment variation can attenuate this signal. This is a statement about the target, not a monotonic-power claim: test power also depends on score variance and the treatment relevance gate. With multiple roots, Equation (10) additionally permits cancellation.

A.7 Standard residual orthogonality and calibration

The uncorrected ablations use the standard frozen-score residual-product guarantee. Fix a score A independently of the test observations. For either comparison, abbreviate its response and conditioning variables by (V_e, X_e), let $\eta_e = \mathbb{E}[V_e \mid X_e]$ and $\rho_{e,j} = \mathbb{E}[h_j(A) \mid X_e]$, and set $\psi_{e,j} = (V_e - \eta_e)\{h_j(A) - \rho_{e,j}\}$. Conditional on development data,

$$\mathbb{E}[\widehat{\psi}_{e,j} - \psi_{e,j}] = \mathbb{E}[(\widehat{\eta}_e - \eta_e)(\widehat{\rho}_{e,j} - \rho_{e,j})]. \tag{A.11}$$

The first-order terms vanish by conditional centering, leaving a product of regression errors (Shah & Peters, 2020; Chernozhukov et al., 2018). This does not account for estimating the anchor itself.

For fixed J, suppose these products are $o_{\mathbb{P}}(n_{\mathrm{test}}^{-1/2})$, the fitted products are L_2 consistent, and true and fitted products have uniformly bounded moments of order $q > 4$. Assume the test observations are i.i.d. and independent of all fits. For the variance-screening rule, also assume population standard deviations stay away from the cutoff 10^{-12} and retained variances are bounded away from zero. The conditional central limit theorem and multiplier approximation then justify the component tests under their *fitted-score* nulls (Chernozhukov et al., 2013). The observed statistic and a fixed number of independent multiplier draws share a continuous limiting null law, making the plus-one Monte Carlo rank conservative. With no retained coordinates the reported value is one.

For a frozen-score union null, at least one component null holds. Hence, for that component e, $\mathbb{P}\{\max(p_T, p_Y) \leq u\} \leq \mathbb{P}(p_e \leq u) \leq u + o(1)$, without requiring independence (Berger & Hsu, 1996). A union bound gives the prespecified-block adjustment in Equation (B.3) when every block-specific null holds.

None of these frozen-score statements by itself transfers the ideal-anchor causal null to a learned score. The uncorrected procedure additionally needs a bridge such as Equation (A.6); the correction in Section 4.2 instead targets the ideal moment directly under product-rate conditions.

B Plug-in implementation and ablations

This appendix specifies the uncorrected ablations. The corrected procedure retains the same representation architectures but uses the fully nested fitting and inference protocol in Appendix G.

B.1 Representations, fitting, and data separation

The full-proxy ablation gives the uncompressed proxy vector $\boldsymbol{W}$ to the treatment regressions. The oracle variant keeps only the simulator-designated treatment-proxy columns $\boldsymbol{W}_T$ and is not available

in an observational application without such design knowledge. PCA retains the first two components of standardized $\boldsymbol{W}$. PLS retains two components fitted to predict T from $\boldsymbol{W}$ and never uses Y. PCA and PLS are fitted on the development sample. The blockwise variant applies the complete test to each of B prespecified proxy groups and reports

$$p_{\text{block}} = \min\{1, B \min_{b \le B} p_{\text{AO}}^{(b)}\}.$$

This controls selection across blocks but requires the groups to be supplied before testing.

All neural models have two 32-unit hidden layers with activation $\text{SiLU}(x) = x/(1 + e^{-x})$ in every encoder, decoder, and prediction head. Their Gaussian encoders return the posterior mean at evaluation. With mean-squared errors on standardized variables, let ℓ_W be proxy reconstruction error and ℓ_T, ℓ_Y be treatment and outcome prediction errors. Let KL denote the Kullback–Leibler divergence of the encoded Gaussian distribution from the standard normal prior $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Then the training objectives are

$$\begin{aligned} \mathcal{L}_{\text{recon}} &= \ell_W + .01\,\text{KL}, \\ \mathcal{L}_{\text{treat-VAE}} &= \ell_W + 1.5\ell_T + .01\,\text{KL}, \\ \mathcal{L}_{\text{VIB}} &= 1.5\ell_T + .01\,\text{KL}. \end{aligned} \tag{B.1}$$

The reconstruction VAE decodes $\boldsymbol{W}$ from a two-dimensional $\boldsymbol{Z}$ and uses neither T nor Y. The treatment VAE both decodes $\boldsymbol{W}$ and predicts T from $(\boldsymbol{Z}, \boldsymbol{C}_T)$; the variational information bottleneck (VIB) has the same two-dimensional encoder and treatment head but no proxy decoder. Neither treatment-focused objective uses Y.

A structured ablation, motivated by TEDVAE (Zhang et al., 2021), learns three one-dimensional blocks (Z_c, Z_t, Z_y). It decodes $\boldsymbol{W}$ from all three, predicts T from $(Z_c, Z_t, \boldsymbol{C}_T)$, and predicts Y from $(Z_c, Z_y, T, \boldsymbol{C}_Y)$. Its loss is $\ell_W + 1.5\ell_T + \ell_Y + .01 \sum_{b \in \{c,t,y\}} \text{KL}_b$. One variant supplies only Z_c to the anchor; another supplies all three coordinates. These are continuous-treatment representation ablations, not the published TEDVAE effect estimator. They may use outcomes in the non-test sample, but never a final-test outcome. Two additional stress-test rows omit the anchor and apply the transformation bank separately to every coordinate of either the reconstruction VAE or all three structured blocks. They are deliberately unanchored diagnostics, not AO-CES variants.

Adam uses learning rate 3×10^{-3}, full batches, at most 60 epochs, and patience 15. After reserving 20% of the observations as the final test set, neural fitting divides the remaining rows into portions containing 60% and 20% of the full sample for training and validation, respectively. Neural standardizers are fitted on the 60% portion; early stopping uses only the 20% validation portion. Non-neural representations use the 80% development sample. No final-test row is used to fit a tested representation or regression, or to select hyperparameters.

For a fixed representation $\boldsymbol{Z}$, the compressed ablations estimate A_Z, not necessarily A_0. Two shuffled folds of the development sample form out-of-fold estimates

$$\widehat{A}_{Z,i} = \widehat{\mathbb{E}}[T \mid \boldsymbol{Z}, \boldsymbol{C}_T] - \widehat{\mathbb{E}}[T \mid \boldsymbol{C}_T].$$

The representation itself is fitted once before this two-fold step; thus the cross-fitting applies to the anchor regressions, not necessarily to a supervised representation learner. Before constructing targets, set $\widehat{A}_{\text{plug},i} = \text{clip}(\widehat{A}_{Z,i}, -\kappa, \kappa)$ with $\kappa = 4$. The clipped out-of-fold anchors supply the $h_j(\widehat{A}_{\text{plug},i})$ targets used to train the nuisance regressions. Both treatment regressions are then refitted on all development rows to construct the final-test anchor. The two multi-output nuisance fits regress $(T, h_1(\widehat{A}_{\text{plug}}), \ldots, h_8(\widehat{A}_{\text{plug}}))$ on $\boldsymbol{C}_T$ and $(Y, h_1(\widehat{A}_{\text{plug}}), \ldots, h_8(\widehat{A}_{\text{plug}}))$ on $(T, \boldsymbol{C}_Y)$, respectively, using all development rows and predicting the final test once.

Every conditional-mean fit uses the same *FeatureRidge* dictionary: standardized inputs, all non-constant monomials through degree two, 96 random Fourier features, and ridge penalty 2 with an intercept. The radial basis function (RBF) bandwidth is the median positive squared distance among at most 256 fitted rows, giving kernel parameter $\gamma_{\text{RBF}} = (2\,\text{median})^{-1}$. The plug-in ablations restrict treatment prediction to the chosen representation and clip the fitted difference:

$$\widehat{A}_{\text{plug}} = \text{clip}\{\widehat{m}_Z(\boldsymbol{Z}, \boldsymbol{C}_T) - \widehat{m}_0(\boldsymbol{C}_T), -\kappa, \kappa\}. \tag{B.2}$$

Here $\kappa = 4$ is a fixed numerical safeguard rather than a causal or theoretical constant. The following transformations are applied after clipping:

$$h(a) \in \{a, a^2, a^3, \tanh a, \sin a, \cos a, \\ \exp[-(a-1)^2/2], \exp[-(a+1)^2/2]\}.$$

B.2 Statistic and multiplier calibration

When scientific knowledge supplies B proxy groups $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(B)}$, run the complete two-test procedure on each. With $p_{\mathrm{AO}}^{(b)}$ denoting the value for group b, report

$$p_{\mathrm{block}} = \min\{1,\ B \min_{b \le B} p_{\mathrm{AO}}^{(b)}\}. \tag{B.3}$$

The Bonferroni factor corrects for searching across the supplied groups, which must be fixed before testing.

For the treatment test $e = T$, set $(V_e, \boldsymbol{X}_e) = (T, \boldsymbol{C}_T)$; for the outcome test $e = Y$, set $(V_e, \boldsymbol{X}_e) = (Y, (T, \boldsymbol{C}_Y))$. On the final test sample,

$$\widehat{\psi}_{e,i,j} = \{V_{e,i} - \widehat{\eta}_e(\boldsymbol{X}_{e,i})\}\{h_j(\widehat{A}_{\mathrm{plug},i}) - \widehat{\rho}_{e,j}(\boldsymbol{X}_{e,i})\}.$$

Let $\overline{\psi}_{e,j}$ and $s_{e,j}$ be its sample mean and sample standard deviation. Coordinates with $s_{e,j} \le 10^{-12}$ are removed, and

$$M_e = \max_j \frac{\sqrt{n_{\mathrm{test}}}|\overline{\psi}_{e,j}|}{s_{e,j}}. \tag{B.4}$$

For each of $R = 499$ replicates, independent Rademacher signs $\xi_i^{(r)}$, equal to -1 or 1 with probability $1/2$ each, give

$$M_e^{(r)} = \max_j \frac{|\sum_i \xi_i^{(r)}(\widehat{\psi}_{e,i,j} - \overline{\psi}_{e,j})|}{\sqrt{n_{\mathrm{test}}}\, s_{e,j}}, \qquad p_e = \frac{1 + \sum_{r=1}^{R} \mathbf{1}\{M_e^{(r)} \ge M_e\}}{R+1}. \tag{B.5}$$

If every coordinate is removed, $p_e = 1$. The smallest attainable value is .002, and the overall value is $p_{\mathrm{AO}} = \max(p_T, p_Y)$ at level .05. The random-sign bootstrap preserves each row's residual magnitude and the dependence among its transformation coordinates; it is not residual permutation or conditional randomization. The indicator $\mathbf{1}\{M_e^{(r)} \ge M_e\}$ equals one when the bootstrap statistic is at least the observed statistic, and zero otherwise.

C Data-generating mechanisms

In the six fully synthetic mechanisms, all exogenous variables are mutually independent standard Gaussian except where the linear-Laplace mechanism is specified. For these mechanisms, $n = 800$, $\boldsymbol{C}_T = C_1$, $\boldsymbol{C}_Y = C_2$, proxy-noise standard deviation is .45, $H \in \{0, 1\}$ selects the null or alternative, and $\gamma = .8$. The outcome retains a causal dependence on T when $H = 0$; the null removes only the latent path to Y that does not pass through T.

Linear Gaussian and linear Laplace. For $j = 1, \ldots, 4$,

$$W_j = (.65 + .12j)U + .45\varepsilon_{W_j}, \quad T = .65U + .55C_1 + .55\varepsilon_T, \quad Y = T + .45C_2 + \gamma HU + .55\varepsilon_Y.$$

All displayed exogenous variables are Gaussian in the first mechanism and mean-zero, variance-one Laplace in the second.

Nonlinear additive. The four proxy means are $\tanh U$, $\sin(1.3U)$, $U + .25U^2$, and $\mathrm{sign}(U)\sqrt{|U|}$, each with independent .45-SD noise, and

$$T = \sin U + .35U^2 + .55C_1 + .50\varepsilon_T, \qquad Y = 1.10 \tanh T + .40C_2 + \gamma H \cos U + .50\varepsilon_Y.$$

Post-nonlinear.

$$W_j = \tanh\{(.8 + .15j)U + .45\varepsilon_{W_j}\},$$
$$T = \tanh(.9U + .5C_1 + .55\varepsilon_T),$$
$$Y = \operatorname{sign}(Y_0)\log(1 + |Y_0|), \quad Y_0 = .9T + .45C_2 + \gamma H \tanh(1.2U) + .45\varepsilon_Y.$$

Independent treatment and outcome factors. For any scalar x, let

$$v(x) = \left(x, \tanh x, \sin(1.3x), \operatorname{sign}(x)\sqrt{|x|}\right) + .45\boldsymbol{\varepsilon}.$$

Here $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_4)$ is redrawn independently for each block. Independent U_T and U_Y generate the first two four-proxy blocks $v(U_T)$ and $v(U_Y)$. Under $H = 0$, the third block is four independent standard-Gaussian noise variables and

$$T = .80\sin U_T + .35U_T + .45C_1 + .55\varepsilon_T, \qquad Y = \tanh T + .45C_2 + .85\cos U_Y + .55\varepsilon_Y.$$

Under $H = 1$, the third block is $v(U_C)$ for an independent U_C, $.65\sin U_C$ is added to T, and $\gamma\{.65U_C + .25\sin(2U_C)\}$ is added to Y. The oracle treatment set is the first block under the null and the first plus third blocks under the alternative; the blockwise method receives all three groups without this oracle designation.

Twenty nonlinear views. The proxies are .45-noise-contaminated $\{\sin(kU), \cos(kU) : k = 1, \ldots, 10\}$. With $L(U) = .70\sin U + .45\cos(2U) + .30\sin(3U)$,

$$T = L(U) + .45C_1 + .50\varepsilon_T, \qquad Y = \tanh T + .45C_2 + \gamma H L(U) + .50\varepsilon_Y.$$

WDBC real-covariate semi-synthetic mechanism. All 569 rows of the Wisconsin Diagnostic Breast Cancer covariate matrix are used (Street et al., 1993; Pedregosa et al., 2011); its diagnosis label is never accessed. Feature roles are fixed by name and disjoint. The withheld index source comprises *worst radius*, *worst perimeter*, *worst area*, *worst compactness*, *worst concavity*, and *worst concave points*. The six size proxies are *mean radius*, *mean perimeter*, *mean area*, *radius error*, *perimeter error*, and *area error*; the six shape proxies are *mean compactness*, *mean concavity*, *mean concave points*, *compactness error*, *concavity error*, and *concave points error*. Finally, C_1 is *mean texture* and C_2 is *mean smoothness*.

Each selected column is standardized. The latent U is the standardized first principal-component score of the six withheld index columns, with its sign fixed to give a positive loading on *worst radius*. This PC is part of the data generator and is unavailable to the tested methods. Independent Gaussian errors are redrawn for each seed and

$$T = .80U + .30\tanh U + .40C_1 + .55\varepsilon_T, \qquad Y = .90\tanh T + .40C_2 + \gamma HU + .55\varepsilon_Y.$$

All 12 observed proxies are oracle treatment-proxy columns; size and shape are the two prespecified blocks. For a fixed seed, the null and alternative share the covariates, treatment, and errors and differ only by the γU term.

D Complete plug-in ablation results

Every entry below is null rejection rate / alternative detection rate at level .05. All synthetic cells use seeds 0–99 separately under $H = 0$ and $H = 1$. WDBC core and reference-method cells also use 100 seeds per state; WDBC neural cells use 50. A dash denotes a combination that was not run.

On the independent-factor null, anchoring the reconstruction representation reduces rejection from .72 to .05; the paired difference is $-.67$ with a 10,000-draw percentile-bootstrap 95% interval $[-.76, -.57]$. For the three-block representation the corresponding change is .70 to .04, or $-.66$ $[-.75, -.57]$. Prespecified blocks improve independent-factor alternative detection from .45 to .98 while changing null rejection from .06 to .02. Across all seven mechanisms, the direct full-$\boldsymbol{W}$ anchor rejects 45/700 null samples and detects 610/700 alternatives. These aggregates do not erase the mechanism-specific warning: the nonlinear-additive null rate is 10/100.

Table 2: Plug-in anchored non-neural representations. Each displayed cell has 100 repetitions per state. Oracle-$\boldsymbol{W}_T$ uses simulator labels; blockwise uses three groups for independent factors, one for the 20-view mechanism, and two for WDBC.

Method	Lin.-G	Lin.-Lap.	Nonlin.-add.	Post-nonlin.	Indep.	20-view	WDBC
Full $\boldsymbol{W}$	.06/1.00	.05/1.00	.10/.65	.07/1.00	.06/.45	.07/1.00	.04/1.00
PCA–2	.04/1.00	.05/1.00	.08/.68	.06/1.00	.05/.65	.06/.84	.05/.97
PLS–2	.06/1.00	.04/1.00	.08/.81	.06/1.00	.03/.63	.09/1.00	.09/.99
Oracle $\boldsymbol{W}_T$	–	–	–	–	.09/.53	.07/1.00	.04/1.00
Prespecified blocks	–	–	–	–	.02/.98	.07/1.00	.06/.98
Coordinatewise	.03/1.00	.04/1.00	.01/.99	.02/1.00	.03/.98	.00/.98	.04/1.00

Table 3: Plug-in neural representation and mixing-stress results. Synthetic entries use 100 repetitions per state. WDBC entries use 50. "TED" denotes the continuous-treatment structured ablation described above, not the published TEDVAE estimator.

Representation	Lin.-G	Lin.-Lap.	Nonlin.-add.	Post-nonlin.	Indep.	20-view	WDBC
Reconstruction VAE $\to A$	.06/1.00	.04/1.00	.12/.63	.08/1.00	.05/.68	.04/.96	.10/.98
Treatment VAE $\to A$	.04/1.00	.03/1.00	.05/.75	.08/1.00	.05/.60	.07/1.00	.12/.98
Treatment VIB $\to A$	.04/1.00	.04/1.00	.07/.81	.07/1.00	.03/.73	.05/1.00	.10/.98
TED common block $\to A$	.05/1.00	.02/1.00	.07/.78	.08/1.00	.08/.71	.10/.96	.06/.98
TED all blocks $\to A$	–	–	–	–	.04/.65	–	.10/.98
Unanchored pooled VAE	–	–	–	–	.72/.97	–	–
Unanchored pooled TED	–	–	–	–	.70/1.00	–	–

D.1 Reference-method audit

HOLS uses the centered low-dimensional formulas of Schultheiss et al. (2024) on the design $(T, \boldsymbol{C})$, where $\boldsymbol{C}$ combines $\boldsymbol{C}_T$ and $\boldsymbol{C}_Y$ without duplicate columns, and reports the raw coefficient-equality p-value for T. KRCD follows Algorithm 1 of Chen et al. (2026): $Z = (T, \boldsymbol{X})$, the first 40 observations define 40 kernel-basis functions, the ridge constant is 10^{-8}, and the 40 coefficient tests receive a Bonferroni correction. KRCD-C sets $\boldsymbol{X} = \boldsymbol{C}$; KRCD-CW sets $\boldsymbol{X} = (\boldsymbol{C}, \boldsymbol{W})$. Both degree-two polynomial and median-bandwidth RBF kernels are shown. These methods retain their native regression-specification targets, which are not the proxy-moment union null tested by AO-CES.

Table 4: HOLS and KRCD native decisions, 100 repetitions per state. Entries are rejection without / with a shared hidden cause on the same generated data; these are not false-positive rate / power for the AO-CES null.

Method	Lin.-G	Lin.-Lap.	Nonlin.-add.	Post-nonlin.	Indep.	20-view	WDBC
HOLS	.06/.03	.07/.30	1.00/1.00	.07/.82	1.00/.92	.97/1.00	.75/.54
KRCD-C poly	.04/.01	.01/.12	.51/.68	.00/.02	.12/.02	.11/.17	.17/.17
KRCD-C RBF	.04/.07	.65/.69	.01/.02	.02/.05	.08/.01	.03/.03	.04/.08
KRCD-CW poly	.00/.01	.04/.01	.02/.03	.03/.07	.12/.11	.09/.13	.08/.12
KRCD-CW RBF	.19/.48	.89/.97	.04/.05	.04/.07	.09/.12	.03/.09	.39/.72

RCD uses the official implementation provided by `lingam==1.11.0` with variable order $(T, Y, \boldsymbol{C})$ or $(T, Y, \boldsymbol{C}, \boldsymbol{W})$. A NaN in either orientation of the T–Y adjacency pair is counted as a latent-confounder decision. RCD does not return a pair-level p-value and cannot represent a bow containing both $T \to Y$ and latent confounding; these generators contain such a direct edge. Because of its cost and linear non-Gaussian scope, it was run only on the linear-Laplace mechanism for 30 seeds per state:

RCD input	Null decisions	Alternative decisions
$(T, Y, \boldsymbol{C})$	21/30 (.70)	15/30 (.50)
$(T, Y, \boldsymbol{C}, \boldsymbol{W})$	20/30 (.67)	5/30 (.17)

Accordingly, Table 4 and the RCD counts are same-data diagnostics, not evidence that one procedure dominates another under a common estimand.

D.2 Published proxy-test scope check

We use the runnable proxy-variable test of Liu et al. (2024) as a published different-target comparison. Its null is $T \perp\!\!\!\perp Y \mid U$: it asks whether a causal T–Y relation remains after accounting for a latent variable through a designated proxy. By contrast, every mechanism above retains the direct $T \to Y$ relation in both AO-CES conditions and changes only whether a hidden factor also affects Y. Both columns are therefore alternatives for the Liu test, and its first column is not a hidden-confounding false-positive rate.

We nevertheless ran the authors' official code (repository commit `0b19e12e`) as a scope check on the two unchanged linear mechanisms. Proxy coordinate W_1 was fixed in advance; following the repository's one-proxy heuristic, T, W_1, and Y were discretized into 16, 10, and 5 quantile bins. All 400 runs completed:

Mechanism	No hidden back-door cause	Hidden back-door cause
Linear Gaussian	100/100	100/100
Linear Laplace	98/100	98/100

The nearly identical columns are consistent with the test reacting to the retained causal relation, as its target requires, rather than distinguishing the two AO-CES conditions. The kernel proxy test of Wu et al. (2025) has the same causal-edge null and additionally needs designated negative-control roles that the unclassified proxy collection does not supply; running it would not resolve the target mismatch. The TEDVAE-motivated rows in Table 3 instead compare a published proxy-representation structure inside the AO-CES anchor, while remaining explicitly distinct from the published TEDVAE effect estimator.

D.3 Denominators, audit, and empirical limitations

Ten frozen campaigns contain 18,820 seed-level rows: 2,400 non-neural standard-mechanism rows; 2,000 non-neural targeted rows; 4,000 neural standard-mechanism rows; 1,400 independent-factor representation stress rows; 1,000 WDBC non-neural rows; 500 WDBC neural rows; 6,000 HOLS/KRCD synthetic rows; 1,000 HOLS/KRCD WDBC rows; 120 RCD rows; and 400 official Liu scope-check rows. Automated checks verify uniqueness of every $(\text{mechanism}, H, \text{method}, \text{seed})$ key within its campaign, finite statistics, p-values in $[0, 1]$, consistency of decisions with $p < .05$ whenever a p-value exists, and explicit retention of any execution error. All 18,820 rows completed and none was recorded as an error. Wilson intervals for individual rates and paired 10,000-draw percentile-bootstrap intervals for prespecified method differences are computed from the seed-level rows.

The experiments are finite-sample diagnostics, not proof of the representation-fidelity condition. Null rejection reaches .10–.12 for some learned representations and .10 for direct $\boldsymbol{W}$ on the nonlinear-additive mechanism. The direct anchor has low power in the independent-factor mechanism; improvement from testing separate blocks requires scientifically credible groups and pays multiplicity correction. Oracle-$\boldsymbol{W}_T$ is not implementable without proxy-role knowledge. The WDBC experiment preserves real covariate geometry but generates both T and Y and is therefore a positive control, not a claim about a natural treatment effect. Finally, HOLS and KRCD provide confounding diagnostics under regression restrictions, not under the proxy-moment assumptions alone; RCD's 30-seed linear-Laplace check is not aggregated with the 100-seed AO-CES results.

E ADDITIONAL PROXY COMPARISONS

The compressed-anchor comparisons in this appendix use the clipped plug-in procedure of Appendix B, not the corrected full-proxy procedure. They isolate the behavior of compression and provide reference points for the matched corrected comparisons in Appendix G.

Coordinatewise testing and rotation. For each W_k, standardize using development rows, apply the eight functions in Appendix B.1, and fit the treatment and outcome residual regressions on development rows only. Each coordinate receives two max-over-function multiplier tests, combined as

$$p_{\text{coord}} = \min\{1, d_W \min_{k \le d_W} \max(p_{T,k}, p_{Y,k})\}. \tag{E.1}$$

Here d_W is the number of proxy coordinates; the same coordinate must supply both associations. This controls the union of coordinatewise moment alternatives if every coordinate's union-null p-value is valid. It does not turn a mixed coordinate into a valid negative control. The original seven mechanisms use the same samples and splits as the anchored runs.

The rotation follow-up was specified after seeing the coordinatewise results. With a 12×12 matrix of independent standard normals generated once using seed 271828, take its QR decomposition and orient the diagonal of R positively. Replace $\boldsymbol{W}$ by $\boldsymbol{W}Q$ in every independent-factor sample. Since Q is orthogonal, the transformation is invertible and preserves $\boldsymbol{W}\boldsymbol{W}^\top$; it does not discard information or change T or Y. No rotation or hyperparameter was selected by performance. This is an exploratory basis-sensitivity check, not a guarantee for arbitrary mixtures. PCA, reconstruction VAE, treatment VAE, and treatment VIB use their unchanged original objectives, two-dimensional representations, and training settings on both rotated conditions and all five binary populations. These 2,800 evaluations share the generated samples and splits used by full-$\boldsymbol{W}$, PLS, and the other comparators. For mixed coordinates, the Wilson 95% intervals for T-VAE are $[.010, .085]$ for null rejection and $[.604, .781]$ for detection. Its .27 detection gain over full-$\boldsymbol{W}$ has paired interval $[.17, .37]$; its .12 gain over PLS has interval $[.05, .19]$. These intervals are descriptive, without multiplicity adjustment or post-selection coverage for a chosen winning representation.

Table 5: Compressed plug-in ablations: null rejection / confounding detection, 100 repetitions per condition. All rows except coordinatewise testing use an AO anchor. These do not use the fitted-anchor correction.

Representation or test	Original coordinates	Mixed coordinates
Full $\boldsymbol{W}$	.06/.45	.06/.43
PCA	.05/.65	.12/.91
PLS	.03/.63	.02/.58
R-VAE	.05/.68	.14/.79
T-VAE	.05/.60	.03/.70
VIB	.03/.73	.03/.56
Coordinatewise	.03/.98	.30/1.00

Binary populations and perturbation test. We use the five populations in Tables 1–2 of Lee (2014). Let U take four values. Its probabilities are $(.25, .25, .25, .25)$ in the confounded cases and $(.30, .25, .25, .20)$ otherwise. In the table below, $e_u = \mathbb{P}(T = 1 \mid U = u)$, $r_u = \mathbb{P}(Y = 1 \mid T = 0, U = u)$, and $b_u = \mathbb{P}(Y = 1 \mid T = 1, U = u)/r_u$. Printed source values are used without attempting to recover unrounded parameters.

Population	(e_1, e_2, e_3, e_4)	(r_1, r_2, r_3, r_4)	(b_1, b_2, b_3, b_4)
Neither	$(.40, .40, .40, .40)$	$(.30, .30, .30, .30)$	$(1.50, 1.50, 1.50, 1.50)$
Outcome only	$(.40, .40, .40, .40)$	$(.40, .30, .25, .20)$	$(1.54, 1.52, 1.45, 1.42)$
Treatment only	$(.80, .60, .40, .20)$	$(.30, .30, .30, .30)$	$(1.50, 1.50, 1.50, 1.50)$
Positive	$(.76, .60, .40, .24)$	$(.6667, .4000, .3333, .1579)$	$(1.2632, 1.1667, 1.2000, 1.0556)$
Negative	$(.24, .40, .60, .76)$	$(.4737, .3333, .2000, .1667)$	$(1.7593, 2.4000, 2.3333, 1.8947)$

Generate $n = 800$ observations and $d_W = 20$ binary proxies. Independently for each coordinate, draw $\mu_k \sim \text{Uniform}(.05, .95)$ and four prevalences $q_{uk} \sim \text{Beta}(19\mu_k, 19(1 - \mu_k))$. Then draw

$W_k \mid U = u \sim \text{Bernoulli}(q_{uk})$, independently across coordinates conditional on U. The beta construction has variance fraction .05. Draw $T \mid U = u \sim \text{Bernoulli}(e_u)$ and $Y \mid T = t, U = u \sim \text{Bernoulli}(r_u b_u^t)$. The contexts are constant. No latent values or prevalences are supplied to the fitted methods. Relative to the source simulations, the sample size and panel size are explicitly fixed here rather than reproduced wholesale.

For proxy k, let n_{kv} count observations with $W_k = v$ and let $\widehat{r}_{tkv}$ be the empirical outcome risk in $(T = t, W_k = v)$. The implementation of Lee's Equations (3) and (5) uses

$$\widehat{\text{RR}}_k = \frac{\sum_{v=0}^{1} n_{kv}\widehat{r}_{1kv}}{\sum_{v=0}^{1} n_{kv}\widehat{r}_{0kv}}, \qquad S_{\text{Lee}} = \left\{ \frac{1}{d_W} \sum_{k=1}^{d_W} \log \widehat{\text{RR}}_k - \log \widehat{\text{RR}}_{\text{crude}} \right\}^2 . \tag{E.2}$$

Here $\widehat{\text{RR}}_{\text{crude}}$ is the ratio of the outcome risks among treated and untreated observations. The square is outside the average. Permute paired (T, Y) rows together against $\boldsymbol{W}$, 499 times, and compute the plus-one upper-tail p-value. Empty treatment-by-proxy cells or undefined log risks, including in a permutation, are recorded as failures without smoothing or replacement. Eight of the 500 full-sample Lee runs fail; all 3,500 accompanying AO/coordinatewise runs complete.

Table 6: Binary proxy populations: compressed plug-in ablations and published diagnostic. AO and coordinatewise (Coord.) entries are rejections out of 100 completed runs. Lee entries are rejections/completions out of 100 attempts; other attempts have undefined statistics. Lee uses all 800 observations; other tests retain the 80/20 split.

Latent role	Lee	Full $\boldsymbol{W}$	PCA	PLS	R-VAE	T-VAE	VIB	Coord.
Neither	3/98	0	0	1	0	0	0	0
Outcome only	14/98	0	0	0	1	0	0	0
Treatment only	20/98	0	0	3	2	1	0	0
Positive confounding	92/99	1	7	5	0	3	4	0
Negative confounding	90/99	0	4	8	3	4	1	0

Low AO detection is not solely a consequence of requiring both tests: full-$\boldsymbol{W}$ passes the treatment branch in only $7/100$ runs in each confounded population, and the outcome branch in $8/100$ positive and $10/100$ negative cases. PLS passes these branches in $23/100$ and $19/100$ positive cases, and $23/100$ and $20/100$ negative cases. Both comparisons are weak at these fixed settings; these counts do not isolate a unique cause of the loss. The same limitation persists with the added representations: T-VAE passes the treatment/outcome tests in $20/12$ of 100 positive runs and $23/13$ of 100 negative runs. Binary outcomes use the same squared-error objectives as the continuous designs, not binary-specific tuning.

A test-sample-only sensitivity check applies the same Lee implementation to the 160 observations reserved for AO's tests. Of 100 attempts per population, the rejection/completion counts are respectively $2/15$, $3/15$, $2/24$, $3/24$, and $5/24$ in the table's order. Sparse-cell failures account for the remaining 398 attempts, preventing a useful power comparison at that sample budget. The sensitivity check was specified after the full-sample results; no original run or setting was replaced.

Audit and uncertainty. The extension contains 7,300 attempted runs: 1,400 coordinatewise original-mechanism runs, 4,000 binary-comparison runs, 1,400 rotated-proxy runs, and 500 test-sample-only Lee runs. Of these, 6,894 complete and 406 have explicitly retained undefined statistics. The 2,800 representation runs all complete, and their 28 smoke-run cells exactly reproduce the corresponding production p-values; smoke runs are not counted as additional evidence. Raw decisions, failure reasons, SHA-256 hashes, completed-denominator Wilson intervals, and 10,000-resample paired intervals are supplied with the code. For example, coordinatewise null rejection of $30/100$ after rotation has Wilson interval $[.219, .396]$, versus $[.028, .125]$ for $6/100$ full-$\boldsymbol{W}$ AO. The five added implementation checks verify direct versus vectorized Lee statistics, paired-permutation batching, deterministic binary populations, same-coordinate combination, and preservation of the proxy Gram matrix.

E.1 WHICH RELATED METHODS CAN BE COMPARED?

Eligibility depends on the scientific question, available inputs, and model restrictions, not solely on whether a paper mentions confounding. HOLS and KRCD share the motivation of diagnosing confounding, but their regression-model assumptions must hold before rejection has that meaning. In particular, nonlinear misspecification can trigger HOLS without a hidden common cause; a universal ranking from such cases would be misleading.

Method family	Role in this comparison
Lee proxy perturbation	Same diagnostic motivation with binary risk ratios; implemented above on its published population designs.
GCM / PCM / weighted GCM	Statistical building blocks, not automatically causal detectors. Coordinate-wise GCM supplies a direct simple baseline; PCM supplies the population treatment projection.
HOLS / KRCD	Implemented with recorded covariates; KRCD also receives proxies as ordinary covariates. Interpretation remains model-dependent.
Liu / Wu proxy tests	Test a remaining causal relation after accounting for latent variables. The direct relation is present under both our labels; Liu official code is a scope check, not a confounding leaderboard.
DANCE / Xie proxy selection	Validate proxy roles under restricted linear latent models, then estimate effects. Their outputs are not calibrated tests of our no-shared-cause null.
CEVAE / TEDVAE / proximal estimators	Estimate effects under latent or bridge-function assumptions. Representation adaptations are labeled as such; effect estimates are not converted into undocumented confounding decisions.
Multiple-environment tests	Require informative heterogeneous environments. Randomly splitting one environment does not supply causal mechanism shifts.
Spectral / IV / trial-based tests	Require, respectively, high-dimensional linear generative restrictions, a valid instrument, or external randomized evidence. These are not supplied by the present generators.
RCD / CAM-UV / graph search	Recover larger graph structure under additional restrictions. The retained direct edge complicates latent-pair mappings; RCD results are reported with this limitation.

The environment-based family includes Karlsson & Krijthe (2023), Mameche et al. (2024), Reddy & Balasubramanian (2024), and Karlsson & Krijthe (2025). The spectral methods of Janzing & Schölkopf (2018); Liu & Chan (2018) use the geometry of a high-dimensional vector of candidate causes, not an unrestricted auxiliary proxy panel for a scalar treatment. Guo et al. (2014) require an instrument, and De Bartolomeis et al. (2024) combine observational and randomized data. The treatment-intent design of Plečko et al. (2026) requires access to treatment decision-makers or corresponding knowledge proxies. These are relevant alternative sources of identification, but creating those inputs would define a new study rather than fill a missing cell of Tables 5–6.

F PARAMETRIC CALIBRATION OF FITTED-ANCHOR CORRECTION

Appendix G reports the complete 12,600-fit matched representation comparison. Here, 2,400 parametric fits isolate the correction's calibration mechanism using identical nuisance predictions, transformations, and test observations for the plug-in and corrected statistics. The grid comprises three sample sizes, two moment conditions, two anchor-error conditions, and 200 seeds per combination.

For a correctly specified calibration check, generate independent standard normal $U_T, U_Y, \epsilon_T, \epsilon_Y$ and set

$$\boldsymbol{W} = (U_T, U_Y), \qquad T = U_T + \epsilon_T, \qquad Y = T + .8HU_T + U_Y + \epsilon_Y, \qquad H \in \{0, 1\}.$$

Contexts are constant, $h(a) = a$, and all nuisance regressions use ordinary least squares. The factors are observed here; this is an idealized test of inference for the outcome moment, not a hidden-factor recovery benchmark. Its target is $\theta_Y(A_0) = .4H$. A second condition adds $N^{-1/4}U_Y$ to each fitted m_1 prediction, including inner-fold predictions, where N is the development-sample size. This uses a known simulator coordinate solely to impose a convergent outcome-only prediction error; the perturbation is not part of the proposed fitted method. For linear h, the squared-anchor-error term in Equation (A.7) is zero: this $O(N^{-1/4})$ contamination still satisfies the weaker product-rate

conditions when the other regressions have the stated least-squares rates. Table 7 reports all 200-seed conditions.

Table 7: Gaussian calibration check: rejection percentages under the zero-moment null and nonzero-moment alternative, with and without controlled fitted-anchor error. The evaluation sample contains 20% of the n observations.

		Null ($H = 0$)		Alternative ($H = 1$)	
n	Added prediction error	Plug-in	Corrected	Plug-in	Corrected
200	None	2.5	5.0	65.5	70.0
200	$N^{-1/4}U_Y$	38.5	5.0	97.5	69.0
800	None	9.5	8.0	99.5	99.0
800	$N^{-1/4}U_Y$	63.0	8.5	100.0	98.5
3200	None	9.5	5.5	100.0	100.0
3200	$N^{-1/4}U_Y$	90.0	5.5	100.0	100.0

With the controlled error, plug-in null rejection grows from 38.5% to 90%, whereas correction gives 5%, 8.5%, and 5.5%. Average null moments change from $.2931, .1946, .1432$ to $.0020, -.0059, .0009$. The remaining 8.5% rejection at $n = 800$ has Wilson 95% interval $[5.4, 13.2]\%$: the guarantee is asymptotic, not exact finite-sample calibration. The experiment demonstrates removal of the first-order drift while retaining high detection of the nonzero moment at larger sample sizes. Raw branch decisions, complete-grid audits, Wilson intervals, and 10,000-resample paired intervals are supplied with the code; the latter are descriptive, without method-selection adjustment.

G Full-proxy representation comparisons

We evaluate the fitted-anchor correction with six treatment predictors under one experimental protocol. The first uses $(\boldsymbol{W}, \boldsymbol{C}_T)$ directly. The others append two features learned by PCA, treatment-supervised PLS, a reconstruction VAE, a treatment-supervised VAE, or a treatment variational information bottleneck. All six retain the original proxy vector $\boldsymbol{W}$; the correction regressions q_j also use $(\boldsymbol{W}, \boldsymbol{C}_T)$. Thus the comparison concerns the predictive value of additional learned features, not replacing $\boldsymbol{W}$ by a compressed representation.

The grid contains all seven original mechanisms in both confounding states, the rotated independent-factor mechanism in both states, and the five binary proxy populations: 21 conditions, six predictors, and 100 seeds per condition, for 12,600 fits. Each fit produces both the corrected outcome statistic and its matched plug-in ablation from the same estimated functions and final-test observations. Every method uses the same 80/20 development/test split, giving 640/160 observations for $n = 800$ and 455/114 for WDBC. Two inner development folds generate out-of-fold targets; each fold refits the representation and first-stage regressions without its held-out rows. Final first-stage regressions are refitted on all development observations. The ridge penalty is 2, with quadratic features and 96 random Fourier features; neural objectives, architecture, early stopping, and the 60-epoch maximum are fixed across conditions.

The identity test uses $h(a) = a$. The separately reported smooth bank applies the eight transformations to $s(a) = 4\tanh(a/4)$ and differentiates the full composition. Outcome tests use two-sided normal tails and standard-deviation floor .01, with Bonferroni correction across the eight smooth coordinates. Treatment gates use 499 multiplier draws, and the final decision rejects when $\max(p_T, p_Y) < .05$. Neither representation nor bank is selected using final-test outcomes. The tables report every condition for both banks and both statistics.

The accompanying code records seed-level branch decisions, empirical moments, failures, Wilson intervals, and paired-bootstrap contrasts. Across all null conditions, corrected final rejection reaches 10% for the identity bank and 8% for smooth8. The outcome test alone reaches 14%: low final rejection in some binary populations partly reflects a weak treatment-relevance gate, not demonstrated outcome-test calibration. The paired contrasts resample whole seeds 10,000 times and are descriptive, without adjustment for choosing a method after observing results. The parametric calibration experiment in Appendix F evaluates the same correction with correctly specified least-squares regres-

Table 8: Harmonized linear comparison. Every method retains full $\boldsymbol{W}$; the five feature-augmented variants add two learned features. Entries give plug-in/corrected rejection percentages, with null and alternative states separated into panels; 100 repetitions per cell.

Mechanism	Full $\boldsymbol{W}$	+PCA	+PLS	+R-VAE	+T-VAE	+VIB
Null rejection (%)						
Gaussian	9/6	8/6	9/7	7/6	8/8	7/6
Laplace	4/4	3/5	4/5	1/3	1/3	2/4
Nonlinear additive	7/5	7/6	6/5	6/6	6/7	6/5
Post-nonlinear	9/8	9/9	8/10	10/9	9/10	8/9
Independent factors	5/7	5/7	5/7	4/4	4/8	3/5
Harmonic proxies	6/8	6/9	6/7	7/8	7/6	8/6
WDBC	4/6	5/6	5/6	5/6	5/6	6/5
Rotated factors	4/8	4/7	4/7	6/5	5/6	4/7
Alternative rejection (%)						
Gaussian	100/100	100/100	100/100	100/100	100/100	100/100
Laplace	100/100	100/100	100/100	100/100	100/100	100/100
Nonlinear additive	15/16	19/17	16/16	20/15	22/15	22/15
Post-nonlinear	100/100	100/100	100/100	100/100	100/100	100/100
Independent factors	48/63	48/63	48/63	44/62	43/66	45/62
Harmonic proxies	100/99	100/99	100/99	100/99	100/99	100/100
WDBC	96/79	95/79	95/78	95/77	93/76	94/76
Rotated factors	48/65	49/64	48/64	47/63	46/63	43/63
Binary populations: rejection (%) in the stated condition						
Neither	0/0	0/0	0/0	0/1	0/0	0/0
Outcome only	0/0	0/0	0/0	0/1	0/0	0/0
Treatment only	0/0	0/0	0/0	0/2	0/1	0/0
Positive confounding	1/3	1/3	0/2	0/2	1/1	0/1
Negative confounding	1/1	1/1	0/1	1/1	1/1	0/1

Table 9: Harmonized smooth8 comparison. Every method retains full $\boldsymbol{W}$; the five feature-augmented variants add two learned features. Entries give plug-in/corrected rejection percentages, with null and alternative states separated into panels; 100 repetitions per cell.

Mechanism	Full $\boldsymbol{W}$	+PCA	+PLS	+R-VAE	+T-VAE	+VIB
Null rejection (%)						
Gaussian	3/3	3/2	4/3	2/2	3/2	2/1
Laplace	2/3	2/2	2/3	2/3	2/3	3/3
Nonlinear additive	4/6	5/5	5/5	4/5	5/4	5/4
Post-nonlinear	2/3	2/2	3/2	2/2	2/2	2/2
Independent factors	1/1	1/1	1/1	4/1	2/2	3/2
Harmonic proxies	3/7	3/7	3/7	4/8	1/7	2/7
WDBC	2/2	3/2	3/3	2/2	2/3	3/2
Rotated factors	1/1	1/2	1/1	2/1	2/1	1/1
Alternative rejection (%)						
Gaussian	100/100	100/100	100/100	100/100	100/100	100/100
Laplace	100/100	100/100	100/100	100/100	100/100	100/100
Nonlinear additive	41/40	44/43	44/39	38/40	38/41	39/41
Post-nonlinear	100/100	100/100	100/100	100/100	100/100	100/100
Independent factors	29/35	28/38	28/38	28/35	30/36	31/39
Harmonic proxies	100/98	100/98	100/98	100/98	100/99	100/100
WDBC	97/78	97/80	96/80	97/79	97/77	97/78
Rotated factors	28/40	28/39	27/39	25/35	29/36	27/36
Binary populations: rejection (%) in the stated condition						
Neither	0/1	0/1	0/1	0/0	0/0	0/0
Outcome only	0/2	0/2	0/2	0/0	0/0	0/0
Treatment only	0/0	0/0	0/0	0/0	0/0	0/0
Positive confounding	0/1	0/1	0/1	0/0	1/0	0/0
Negative confounding	0/2	0/2	0/2	0/2	0/1	0/1

sions; the representation comparisons do not themselves establish the regression-rate assumptions of Theorem 1.

H INDEPENDENT FCI SCREENING AND PROXY ANNOTATIONS

The application uses four treatment–outcome modules and two recorded baseline covariates. For $i = 1, \ldots, 4$, generate

$$\begin{aligned} T_i &= .6C_1 + .4C_2 + U_i + \varepsilon_{T_i}, \\ Y_i &= .7T_i + .3C_1 - .3C_2 + .8H_iU_i + V_i + \varepsilon_{Y_i}, \\ W^T_{i,r} &= U_i + .5\eta^T_{i,r}, \qquad W^Y_{i,r} = V_i + .5\eta^Y_{i,r}, \quad r = 1, 2. \end{aligned} \tag{H.1}$$

All roots and errors are mutually independent standard Gaussians. The fixed indicator vector H is $(0, 0, 0, 0)$, $(1, 1, 0, 0)$, or $(1, 1, 1, 1)$, yielding zero, two, or four confounded direct relations. All sixteen proxy measurements are supplied together, without their root labels. Each query $T_i \to Y_j$ uses $\boldsymbol{C}_T = \boldsymbol{C}_Y = (C_1, C_2)$. These baseline contexts are prespecified; no oracle parent sets or latent variables are supplied. Treatment/outcome roles specify the query direction, not a new orientation inferred from its test.

FCI from causal-learn 0.1.4.8 (Zheng et al., 2024) uses an independent 25% of rows, the ten primary variables, Fisher-z tests at .05, unlimited search depth/path length, and no supplied background knowledge. Its cross-tier adjacencies select among sixteen eligible queries. The remaining rows are split 80/20 for development and testing: 60% and 15% of the total sample. The full-$\boldsymbol{W}$ OLS implementation uses the identical corrected score, two-fold nuisance fitting, identity transform, fixed standard-deviation floor, and treatment gate as the parametric experiment. Thus Corollary A.1 applies. Plug-in and corrected tests share every fitted prediction. Holm correction at .05 is applied either to the frozen FCI family or to all sixteen queries. No PAG endpoint is modified.

Table 10: FCI's unchanged adjacency F1 against this design's 20-edge latent-projection skeleton, and cross-tier candidate F1. Each row uses 100 independent replicates; brackets give graph-bootstrap 95% intervals. The annotation layer does not repair the missing primary-variable adjacencies.

n	Confounded pairs	Ten-node adjacency F1	Cross-tier F1
1000	0	.739 [.732,.745]	.994 [.988,.999]
1000	2	.721 [.714,.728]	1.000 [1,1]
1000	4	.692 [.685,.699]	1.000 [1,1]
3000	0	.858 [.853,.863]	1.000 [1,1]
3000	2	.817 [.812,.821]	1.000 [1,1]
3000	4	.785 [.779,.791]	1.000 [1,1]

Table 11: Confounding annotations: pooled precision/recall and dataset-wise false-annotation probability, all in percent. A dash denotes undefined recall when no pair is confounded. "All" tests all sixteen queries; "FCI" tests only selected adjacencies. Every entry uses the same 100 replicates as Table 10.

		FCI candidates		All queries	
n	Confounded	Plug-in	Corrected	Plug-in	Corrected
1000	0	0/–/13	0/–/4	0/–/15	0/–/4
1000	2	95.3/90.5/9	98.4/93.0/3	93.1/81.0/12	96.0/83.0/7
1000	4	100/95.0/0	100/97.0/0	95.3/80.8/16	98.2/83.5/6
3000	0	0/–/11	0/–/3	0/–/8	0/–/7
3000	2	94.8/100/11	97.6/100/5	95.7/100/9	96.6/100/7
3000	4	100/100/0	100/100/0	95.9/100/15	97.8/100/8

The complete grid contains 600 dataset configurations and 9,600 matched query fits, with no failures. Equal seeds share simulated roots across confounding states; confidence intervals resample independent replicates within each cell. FCI retained all four true cross-tier adjacencies, so annotation recall conditional on selection equals end-to-end recall here. This favorable screen does not measure robustness to missed target pairs. At $n = 1000$, corrected recall is .930 (graph-bootstrap 95% interval $[.890, .965]$) with two confounded pairs and .970 $[.950, .988]$ with four. Under the all-null model,

corrected FCI-family false-annotation rates are .04 and .03, with Wilson intervals $[.016, .098]$ and $[.010, .085]$. The all-query alternative reduces power at $n = 1000$ and reaches false-annotation rates up to .08. These finite-sample results support adding tested confounding evidence, not improved PAG orientation or unrestricted mixed-graph recovery. Seed-level endpoint matrices, p-values, confidence intervals, and source hashes accompany the code.

H.1 Validity of the graph annotations

The graph experiment uses the same prespecified contexts $\boldsymbol{C}_T = \boldsymbol{C}_Y = (C_1, C_2)$ for all 16 eligible pairs (T_i, Y_j). Write $S_i = T_i - .6C_1 - .4C_2 = U_i + \varepsilon_{T_i}$, $M_i = \mathbb{E}[U_i \mid \boldsymbol{W}]$, and $\lambda_i = \mathrm{Var}(M_i)$. Independent Gaussian roots give $A_{0,i} = M_i$ and $\mathbb{E}[M_i \mid S_i] = \lambda_i S_i/2$. For a diagonal pair, the adjusted outcome is

$$R_{Y_i} = .8H_i(U_i - S_i/2) + V_i + \varepsilon_{Y_i}.$$

Consequently, for the linear transformation,

$$\theta_{Y,1}^{(i,j)} = .4H_i\lambda_i\,\mathbf{1}\{i = j\}. \tag{H.2}$$

Thus every off-diagonal pair and every unconfounded diagonal pair satisfies the ideal outcome null. The two unit-loading views with independent noise variance $1/4$ give $\lambda_i = 1 - (1+4+4)^{-1} = 8/9$, so each confounded diagonal has moment $16/45 > 0$. All correction regressions are affine, satisfying Corollary A.1.

Lemma H.1 (Selected-edge error control)**.** *Fix a finite family of eligible pairs and their pair-specific corrected tests satisfying Theorem 1's conditions, or its product-rate version in Appendix A.5. Use development and test data independent of discovery data, which may select any subset of this family. Applying Holm's procedure at level α to the selected p-values gives asymptotic familywise error at most α for annotations whose ideal outcome null holds.*

Proof. Conditional on the selected subset, its null p-values remain asymptotically super-uniform because discovery uses independent observations. If $m_0 > 0$ selected nulls include a false rejection, Holm's first rejected true null has $p \le \alpha/m_0$. A union bound gives probability at most $\alpha + o(1)$, without independence between edge tests (Holm, 1979). There are finitely many possible subsets, so averaging over discovery preserves the bound. □

This controls false annotations, not discovery errors or PAG endpoint marks; the latter remain unchanged.

I Raw measurement geometry: a matched Gaussian check

This experiment checks Proposition 1 with known population targets. Independent standard Gaussian roots generate

$$T = U_T + \varepsilon_T, \qquad Y = T + .8HU_T + U_Y + \varepsilon_Y, \tag{I.1}$$

where $H \in \{0, 1\}$ switches the shared cause off and on. All measurement errors precede treatment. Let $\boldsymbol{U} = (U_T, U_Y)^\top$, $\boldsymbol{\eta} \sim N(0, I_2)$ be independent sensor noise, and $R = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. The four observed vectors are

$$\underbrace{\boldsymbol{U} + .5\boldsymbol{\eta}}_{\text{separate channels}}, \qquad \underbrace{R(\boldsymbol{U} + .5\boldsymbol{\eta})}_{\text{invertible observed mix}}, \qquad \underbrace{U_T + U_Y + .5\eta_1}_{\text{noisy sum}}, \qquad \underbrace{R\boldsymbol{U} + .5\boldsymbol{\eta}}_{\text{full-rank noisy mix}}.$$

The first two contain exactly the same information; the last two can mix independent causes in the ideal prediction itself. In this order, the unconfounded population moments are $(0, 0, 4/9, 4/29)$ and the confounded moments are $(8/25, 8/25, 28/45, 12/29)$. Every design satisfies proxy mean exclusion, but the last two violate the ideal-anchor outcome null when $H = 0$.

The frozen grid uses 100 common seeds for each design, state, and total sample size $n \in \{800, 3200\}$: 1,600 fitted cells, with no failures. All fits retain the complete observed vector and use the existing OLS identity procedure, an 80/20 development/test split, and two inner folds. Thus the evaluation samples have 160 and 640 observations. On the same evaluation rows we also compute plugin and corrected scores using exact population nuisance functions, labelled oracle benchmarks. All four

variants use the same Gaussian outcome calibration, .01 SD floor, and 499-draw treatment relevance gate as the fitted procedure. The oracle benchmarks use known generating parameters, not fitted test outcomes.

Table 12: Measurement-geometry check. Each rejection entry is a percentage of 100 seeds at nominal .05; P/C denote plugin/corrected scores. The moment column is the mean fitted corrected score. †: when $H = 0$, these measurements violate the ideal outcome null despite no hidden common cause; their rejections are not nominal type-I errors for that statistical null.

Measurements	H	n	Target	Moment	Oracle		Fitted	
					P	C	P	C
Separate	0	800	.000	−.008	7	3	10	6
Separate	0	3200	.000	.001	2	2	6	2
Separate	1	800	.320	.309	97	92	96	92
Separate	1	3200	.320	.322	100	100	100	100
Observed mix	0	800	.000	−.008	7	3	10	6
Observed mix	0	3200	.000	.001	2	2	6	2
Observed mix	1	800	.320	.309	97	92	96	92
Observed mix	1	3200	.320	.322	100	100	100	100
Noisy sum†	0	800	.444	.443	100	100	100	100
Noisy sum†	0	3200	.444	.441	100	100	100	100
Noisy sum†	1	800	.622	.618	100	100	100	100
Noisy sum†	1	3200	.622	.621	100	100	100	100
Full-rank mix†	0	800	.138	.130	46	35	40	34
Full-rank mix†	0	3200	.138	.137	97	93	93	93
Full-rank mix†	1	800	.414	.404	100	100	100	99
Full-rank mix†	1	3200	.414	.413	100	100	100	100

The separate-channel corrected rejection rates without a shared cause are .06 (Wilson 95% interval $[.028, .125]$) and .02 ($[.006, .070]$). Every paired oracle and fitted result is preserved under the invertible observed transformation, within 9.9×10^{-14} numerical error across 400 paired cells; these equal rows are not independent replication. In contrast, the full-rank noisy mix gives corrected rejection rates .34 ($[.255, .437]$) and .93 ($[.863, .966]$) without a shared cause. The noisy sum gives 1.00 ($[.963, 1.000]$) at both sizes. Oracle results show the same failure of the causal measurement restriction. Correction addresses fitted-anchor error; it cannot change a nonzero ideal target.

The raw archive and summary report all four score means, Monte Carlo standard errors, Wilson intervals, and whole-seed paired bootstrap comparisons. For the separate-channel null, the fitted corrected-minus-plugin rejection differences are $-.04$ at both sizes, with paired 95% intervals $[-.11, .03]$ and $[-.10, .01]$; these comparisons do not establish a difference in rejection rates. The 32 smoke cells reproduce exactly, and all scientific classifications and population moments pass the audit.

J Diagnosing binary power with an observed-proxy oracle

The small binary rejection rates in Appendix G can arise from weak available signal, learning error, or the treatment gate. We separate these possibilities in the same five binary populations, using $n \in \{800, 3200\}$ and 100 paired seeds per population and sample size. All arms use full $\boldsymbol{W}$, the identity transform, the original 80/20 development/test split, two inner development folds, the unchanged corrected score and treatment gate, and standard-deviation floor .01. Thus the test sizes are 160 and 640.

Population calculations and matched arms. The binary model has four latent classes, conditionally independent Bernoulli proxies, treatment probabilities e_u, and outcome probabilities r_{tu}. Known simulation parameters give $\pi_u(\boldsymbol{w}) = P(U = u \mid \boldsymbol{W} = \boldsymbol{w})$; realized latent classes are never supplied to a prediction method. Let $m(\boldsymbol{w}) = \sum_u \pi_u(\boldsymbol{w}) e_u$, $b = E[T]$, $g_t = E[Y \mid T = t]$, and

$v = \mathrm{Var}\{m(\boldsymbol{W})\}$. For the identity transform,

$$q(\boldsymbol{w}) = \sum_u \pi_u(\boldsymbol{w})\{(1-e_u)(r_{0u}-g_0) + e_u(r_{1u}-g_1)\},$$
$$r(1) = v/b, \qquad r(0) = -v/(1-b), \qquad k = 0.$$

Enumerating all 2^{20} proxy patterns in blocks gives the population moment $E[q(\boldsymbol{W})m(\boldsymbol{W})]$, v, and the corrected-score variance. The *full oracle* supplies these population functions, conditional on the same observed proxy information. The *oracle-anchor* arm supplies only m, b, retaining the existing nested regressions for the other functions. The *current* arm calls the original full-proxy implementation. This is a diagnostic comparison, not an implementable oracle baseline or a universal upper bound on power.

Table 13: Binary diagnostic: treatment-gate/outcome/final rejection percentages for the corrected identity test, each from 100 paired seeds. All five populations and both sample sizes are shown. "Oracle anchor" supplies only m, b; "Full oracle" supplies all population functions. The logistic follow-up was triggered by the prespecified oracle–current power gap.

Population	n	Current	Oracle anchor	Full oracle	Logistic
Neither	800	2/10/0	0/6/0	0/0/0	6/5/1
Outcome only	800	2/9/0	0/6/0	0/9/0	6/4/1
Treatment only	800	6/10/0	76/4/4	76/3/1	41/4/1
Positive confounding	800	8/9/3	58/6/5	57/30/27	34/8/2
Negative confounding	800	8/7/1	58/6/3	57/36/32	34/8/2
Neither	3200	6/6/1	0/6/0	0/0/0	5/7/1
Outcome only	3200	6/5/0	0/6/0	0/4/0	5/7/0
Treatment only	3200	86/9/8	100/5/5	100/4/4	99/4/3
Positive confounding	3200	68/50/37	98/61/61	98/92/92	97/54/53
Negative confounding	3200	68/40/29	98/53/53	98/83/83	97/47/47

What limits power. At $n = 800$, the full oracle detects 27% of positive and 32% of negative alternatives (Wilson 95% intervals $[19.3, 36.4]$ and $[23.7, 41.7]$). Its treatment gate rejects in 57% of either alternative, and its outcome test rejects in 30% and 36%. The small held-out sample therefore limits even this population-function benchmark. Learning adds substantial noise: the current anchor's mean squared error is .1538, compared with mean true anchor variance .00882. Supplying the true anchor raises final detection only to 5% and 3%; its mean empirical corrected-score standard deviations remain .199 and .195, versus .058 and .054 with all population functions. At $n = 3200$, full-oracle detection reaches 92% and 83% ($[85.0, 95.9]$ and $[74.5, 89.1]$), whereas current detection is 37% and 29% and oracle-anchor detection is 61% and 53%. These comparisons implicate both anchor learning and downstream regression, in addition to sample size. The score floor never binds in any alternative in any arm.

Prespecified probability-learner follow-up. The frozen protocol calls for one follow-up if full-oracle final rejection exceeds current rejection by at least 20 percentage points in any alternative cell. All four cells meet this trigger. The follow-up uses L2 logistic regression on the 20 raw proxy columns for m (fixed $C = 1$), the empirical treatment mean for b, and empirical binary-treatment outcome means for g. The original nested regression procedure for the remaining signed targets, held-out observations, score, and calibration are unchanged. Across the alternatives, anchor error falls to .00828 at $n = 800$ and .00309 at $n = 3200$, but small-sample final detection remains 2%. At $n = 3200$, detection rises to 53% and 47%, compared with 37% and 29% for the current learner: paired differences of 16 and 18 percentage points (descriptive seed-bootstrap 95% intervals $[7, 25]$ and $[9, 28]$). The six follow-up null cells have final rejection between 0% and 3%. These results support a partial practical improvement at the larger sample size, not a resolution of the small-sample binary limitation or verification of nonparametric regression-rate assumptions.

Audit. The primary and follow-up grids contain 3000 and 1000 fitted/evaluated arms, respectively, with zero failures. Eight targeted unit tests and a disjoint 90-arm smoke grid pass. Enumerated probability mass differs from one by at most 3.4×10^{-15}, and the zero-mean and equivalent-moment

identities agree within 5.4×10^{-16}. All 500 current-arm $n = 800$ comparisons reproduce the archived p-values within 1.3×10^{-13} and preserve every decision. Saved results include both plug-in and corrected branch p-values, all population parameters, split and source hashes, prediction errors, score moments, floor indicators, Wilson intervals, and paired seed-bootstrap contrasts. Shared seeds induce paired comparisons within each cell; proxy parameters are regenerated by the original generator at each sample size.

K Randomized graph and sensor-noise stress test

We randomize direct effects and hidden common causes separately to include four kinds of treatment–outcome pair: a direct effect only, a hidden common cause only, both, or neither. There are four treatments T_i, four outcomes Y_j, and two observed baseline variables $\boldsymbol{C} = (C_1, C_2)$. Independent standard-normal variables $U_i, V_j, \epsilon_{T_i}, \epsilon_{Y_j}, C_1, C_2$ generate

$$T_i = .6C_1 + .4C_2 + U_i + \epsilon_{T_i}, \qquad Y_j = \sum_{i=1}^{4} D_{ij}T_i + .3C_1 - .3C_2 + \sum_{i=1}^{4} H_{ij}U_i + V_j + \epsilon_{Y_j}. \quad \text{(K.1)}$$

Here D_{ij} is the direct-effect coefficient and H_{ij} is the coefficient of the unmeasured common cause U_i in the equation for Y_j. Their supports are independent Bernoulli$(.25)$ draws. Each nonzero coefficient independently has a random sign and a magnitude drawn, with equal probability, from $[.15, .30]$ or $[.60, 1.00]$, uniformly within the chosen interval. The all-null condition sets every $H_{ij} = 0$ while retaining the same direct-effect matrix for that seed. The mixed condition retains every drawn coefficient, without requiring each graph to contain a hidden relation.

Every U_i and V_j has two proxy views, each adding independent normal noise of standard deviation σ. Paired conditions use $\sigma = .5$ and 1.5, with identical graph coefficients, primary observations, and standardized sensor errors. The analyst receives all 16 proxy columns, not their latent grouping or the generating parents, and uses $\boldsymbol{C}_T = \boldsymbol{C}_Y = \boldsymbol{C}$ for every query. The ideal identity moment is

$$\theta_{Y,1}^{(i,j)}(A_0) = \tfrac{1}{2}H_{ij}\lambda, \qquad \lambda = \frac{2}{2+\sigma^2}.$$

Thus changing D_{ij} does not change whether the diagnostic's population moment is zero, whereas noisier proxies reduce the magnitude of a nonzero moment. Adjusting for T_i removes its direct contribution to Y_j; the other treatments contribute residual variance but are independent of its anchor given $\boldsymbol{C}$. Conditional on the graph coefficients, every nuisance regression is affine with a nonsingular fixed-dimensional design, so Corollary A.1 applies. A hidden-only pair is a positive case even though its direct coefficient is zero; an annotation does not assert a direct causal arrow.

We generate 100 all-null and 100 mixed graphs with $n = 1000$ observations, each at both noise levels: 400 datasets and 6400 fitted queries. The independent discovery/development/test split is 25/60/15%. FCI with Fisher-z tests at .05 sees only the ten primary variables on discovery observations. Cross-tier adjacencies select candidate pairs; all endpoint marks are preserved. Both inference variants use the unchanged full-proxy, least-squares, identity-only statistic, with Holm correction over either the selected family or all 16 queries. Python hash order is fixed to make the implementation's set traversal reproducible across workers. All 400 datasets completed, and the 12 fixed-protocol smoke datasets reproduced exactly.

FCI selects 200 of 409 confounded pairs: 120 of 303 hidden-only pairs and 80 of 106 pairs with both effects. This selection step therefore limits attainable end-to-end recall before the proxy test is applied. With lower sensor noise, correction changes selected-family recall by $+2.7$ percentage points (paired 95% interval $[0, 5.3]$). With higher noise, the change is -5.9 points ($[-9.4, -2.2]$), while all-null graph error falls from 6/100 to 1/100. The corrected high-noise recall is 10.0% ($[7.2, 13.0]$); the 1/100 null-error rate has Wilson interval $[.2, 5.4]\%$. Correction controls a source of bias, but cannot recover relations omitted by discovery or guarantee high power with weak proxy measurements. Testing all pairs avoids candidate omissions, at the cost of a larger multiplicity penalty; it is not uniformly preferable in Table 14.

Table 14: Complete randomized-graph comparison (%). Error columns give graphs with any false annotation out of 100; mixed-graph precision and recall pool all 409 true hidden relations, including pairs omitted by FCI. The candidate recall is 48.9% for FCI and 100% when all queries are tested.

Sensor SD	Family/score	Null error	Mixed error	Precision	Recall	Cond. recall
.5	FCI/plug-in	4	2	98.4	30.6	62.5
.5	FCI/corrected	3	1	99.3	33.3	68.0
.5	All/plug-in	10	8	93.6	32.3	32.3
.5	All/corrected	3	5	96.8	37.2	37.2
1.5	FCI/plug-in	6	5	92.9	15.9	32.5
1.5	FCI/corrected	1	1	97.6	10.0	20.5
1.5	All/plug-in	3	6	89.5	12.5	12.5
1.5	All/corrected	5	4	87.1	6.6	6.6

Table 15: Annotation counts by relation class in the 100 mixed graphs. P/C denotes plug-in/corrected counts; the same graph draws and FCI candidates are used at both sensor-noise levels. No hidden-only annotation is interpreted as evidence for a direct effect.

			FCI P/C		All P/C	
Relation	Total	Selected	$\sigma = .5$	$\sigma = 1.5$	$\sigma = .5$	$\sigma = 1.5$
Neither	886	1	0/0	0/0	8/4	6/4
Direct only	305	210	2/1	5/1	1/1	0/0
Hidden only	303	120	92/98	50/31	100/111	39/19
Both	106	80	33/38	15/10	32/41	12/8

Uncertainty intervals resample whole graphs 10,000 times, preserving method and sensor pairing; they are descriptive rather than selection-adjusted. The code reports the corresponding breakdown for all-null graphs, every raw branch decision, and full PAG matrices. We do not compare native graph skeletons with the generating directed graph: shared hidden causes can induce additional adjacencies among outcomes in the latent projection.

L WHAT SCREENING AND PROXY ANNOTATIONS ADD

We analyze the saved corrected identity-test results from all 400 randomized datasets in Appendix K, without refitting or changing the PAGs. This is a retrospective descriptive comparison. For each graph, the same query p-values enter Holm's procedure over either the FCI-selected pairs or all 16 pairs. Table 16 separates detections shared by both procedures, detections gained through the smaller selected family, and detections lost because their pairs were not selected.

Table 16: True-confounded queries in 100 mixed graphs per sensor condition; each row totals 409. Gains and losses are relative to testing all pairs. The final column is selected-family minus all-query recall in percentage points, with a paired 95% whole-graph bootstrap interval.

Sensor SD	Both	Gain	Loss	Neither	Recall difference [95% interval]
.5	112	24	40	233	-3.91 $[-7.85,\ 0.00]$
1.5	22	19	5	363	$+3.42$ $[0.79,\ 6.16]$

Restriction cannot remove a Holm rejection on a retained pair. If that pair has rank r among m full-family tests and rank j among k retained tests, then $k-j+1 \leq m-r+1$; its restricted threshold is no smaller, and the same argument applies to every retained predecessor. All 45 lost true detections in Table 16 therefore concern pairs omitted by FCI, whereas all 43 gains concern retained pairs that benefit from reduced multiplicity. The lower-noise gain and loss proportions are 5.87% $[3.54, 8.39]$ and 9.78% $[6.74, 12.93]$; the higher-noise proportions are 4.65% $[2.48, 7.12]$ and 1.22% $[.25, 2.34]$. Intervals use 10,000 resamples of whole graphs, keeping both decisions paired and recomputing the number of true queries within each resample.

The analogous false-finding counts, selected-only/all-query-only, are 1/5 and 1/4 among the 1191 null queries in the mixed graphs, and 3/3 and 1/5 among the 1600 queries in the all-null graphs, at lower and higher noise respectively. No false finding is shared by the two families. Screening therefore changes which null queries can be rejected, not only the total number of rejections.

We also classify the saved endpoint marks, without treating an estimated PAG as ground truth. Of the 136 lower-noise true annotations, 69 have a circle-containing pattern, 27 a directed pattern in either direction, and 40 a bidirected pattern. The corresponding higher-noise counts are 22, 6, and 13; the single false annotation in the mixed graphs at each noise level has a circle endpoint. Directed marks alone do not establish absence of hidden confounding; visibility is additional whole-PAG information (Zhang, 2008). These counts describe evidence added using proxy measurements, not confounders that FCI incorrectly declared absent.

For an illustrative query, choose the first seed and pair with a lower-noise true annotation and the pattern $T_i \circ\!\!\rightarrow Y_j$. This gives seed 14, $T_3 \circ\!\!\rightarrow Y_1$: the generating direct and hidden coefficients are $-.921$ and $-.746$, and the corrected conjunction p-value is .002. Both the three-pair selected family and the 16-pair full family reject. The proxy annotation supplies evidence about a common cause while the circle remains unchanged; it does not orient the PAG. This outcome-selected example is illustrative, not representative.