ACCEPTANCE-SIDE MULTIPLE TESTING FOR PROXY DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Proximal causal learning (PCL) identifies causal effects under unmeasured confounding from valid proxies of the confounder. When proxies must be discovered from high-dimensional data, thousands to millions of candidate proxy sets are screened at once, and the consequential error is *false validation*: retaining an invalid candidate that then biases the causal estimate. We formulate proxy discovery as an acceptance-side multiple-testing problem. A joint tetrad test gives one p-value per candidate triplet, a two-group model converts these into posterior validity scores $L(p) = \mathbb{P}(H = 0 \mid P = p)$, and accepting $L(p) \geq \tau$ has oracle false validation rate (FVR) at most $1 - \tau$. The plug-in rule inherits no guarantee: where invalid p-values retain mass near one, the upward bias of Storey's $\widehat{\pi}_0$, conservative for the FDR, is anti-conservative for the FVR, and the resulting contamination is not identifiable from the p-value distribution alone. An exact identity equates the accepted set's mean absolute bias to its empirical FVR times the mean bias of the accepted invalid triplets. Simulations show where each layer holds and fails; we illustrate the pipeline on 5.7 million candidate triplets from TCGA–BRCA.

1 INTRODUCTION

Unmeasured confounding is a central obstacle to causal inference from observational data; proximal causal learning (PCL; Miao et al., 2018; Tchetgen Tchetgen et al., 2020; Cui et al., 2024) identifies causal effects without observing the confounder U by using a negative-control exposure Z and a negative-control outcome W: two proxies that carry information about U while satisfying exclusion restrictions relative to the treatment T and the outcome O (Appendix A). Standard PCL theory takes suitable proxies as given; in high-dimensional data such as gene expression they must be discovered among the measured variables. The tool is old: in Spearman's one-factor model (Spearman, 1904), variables that depend on one another only through a common factor have rank-one cross-covariances, so certain differences of covariance products, the *tetrads*, vanish. DANCE (Kummerfeld et al., 2024) screens triplets of measured variables, in our application three genes: if the three, together with T and O, depend on one another only through a common latent U, a *disconnected negative-control triplet* (DNCT; Appendix B.1), then six tetrads vanish, any two members can serve as a PCL proxy pair, and DANCE retains the triplets whose tetrads are not rejected.

This test is oriented the opposite way from a conventional hypothesis test, and we inherit that orientation: the null hypothesis is the structure we hope to find, and a triplet is validated and passed to PCL when the test fails to reject. Rejecting a valid triplet only loses a proxy; retaining an invalid one, the classical type II error, is the consequential mistake, because PCL then uses it and the causal estimate is biased. We call this event *false validation*. At scale, with thousands to millions of overlapping candidate triplets, it becomes a multiple-testing problem on the acceptance side. For the set $\mathcal{A}$ of accepted triplets we define the false validation rate $\mathrm{FVR}(\mathcal{A}) = \mathbb{P}(H_i = 1 \mid i \in \mathcal{A})$, where $H_i = 1$ marks an invalid triplet: the contamination of the accepted set, which under a two-group model is the positive false nondiscovery rate of the acceptance region (Storey, 2003). Classical corrections bound errors among rejections and say nothing about it, and a fixed raw-p rule such as "keep $P_i > 0.05$" does not control it either, because its FVR depends on the prevalence π_0 of valid triplets and on how far invalid p-values separate from the null (Figure 2E). We therefore formulate proxy discovery as an acceptance-side multiple-testing problem and carry its error through to the causal estimate.

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.

Contributions. (i) *One test per candidate.* We replace DANCE's six separate tetrad decisions with one candidate-level p-value from the joint Wald statistic of Bollen (1990), which accounts for the covariance among a triplet's four independent tetrads (Section 2.1). (ii) *Acceptance-side validation.* A two-group model (Storey, 2002; Storey & Tibshirani, 2003) converts the p-values into posterior validity scores $L(p) = \mathbb{P}(H = 0 \mid P = p)$, the probability that a triplet with p-value p is valid. Accepting $L(P_i) \geq \tau$ has oracle FVR at most $1 - \tau$ (Proposition 1), but the estimated rule inherits no guarantee, and we identify why: the estimate rests on Storey's tail estimator $\widehat{\pi}_0$, which estimates the proportion of valid triplets from the number of p-values near one, where valid triplets place theirs. Invalid triplets with weak violations also produce p-values near one, so $\widehat{\pi}_0$ overestimates the proportion of valid triplets, and since every score $\widehat{L}(p)$ is proportional to $\widehat{\pi}_0$, every score is inflated and the rule accepts more triplets than its target allows. For the FDR this overestimate is harmless; on the acceptance side it is anti-conservative, and the resulting contamination is not identifiable from the p-value distribution alone, whatever rule is applied to the p-values, a fixed raw-p cutoff included (Proposition 2). What the test can and cannot see, tetrad faithfulness and relevance, is set out in Section 2.1. (iii) *From validation error to causal error.* For every realized accepted set, the mean absolute population PCL bias equals the empirical FVR times the mean absolute bias of the accepted invalid triplets, which separates how often invalid triplets are accepted from how much bias each accepted invalid triplet carries, and bounds the bias of the PCL estimate averaged over the accepted set (Section 2.3). (iv) *Evaluation.* Simulations with known structure test each stage: null calibration, recovery of π_0, FVR and coverage of the plug-in rule, transport of one threshold where a raw-p cutoff must be re-tuned, loss of information under dependence, and the gap between the structural bias of the accepted set and the finite-sample PCL error (Section 3). We illustrate the pipeline on TCGA–BRCA expression data (Section 4).

Related work. Standard PCL assumes suitable proxies are given (Miao et al., 2018; Tchetgen Tchetgen et al., 2020; Cui et al., 2024). Proxy discovery has been approached through the constraints that a valid proxy set implies: DANCE (Kummerfeld et al., 2024) uses vanishing tetrads in a linear one-confounder model, Xie et al. (2024) use rank or independence constraints under several confounders, Ying et al. (2025) extend tetrads to nonlinear latent relations, and pure-measurement-model search (Silva et al., 2006) rests on the same tetrad constraints. In each of these a candidate set is accepted when fixed-level tests of its constraints do not reject, and none controls what fraction of the accepted sets is invalid; our acceptance layer needs only a calibrated candidate-level p-value, so it is placed on top of any of these tests rather than replacing them (Appendix E.4). On the multiple-testing side the error we control is not new: the proportion of false non-rejections is the FNR (Genovese & Wasserman, 2002), its positive version the pFNR (Storey, 2003; Sarkar, 2006), and under the two-group model (Storey, 2002; Efron, 2010) $\mathrm{FVR}(\mathcal{A})$ is the pFNR of the acceptance region (Storey, 2003) and L the local false discovery rate (Efron et al., 2001). Three things are specific to proxy validation. Each accepted triplet is used individually in PCL, so we threshold the candidate-level score L rather than a tail-area quantity that describes a set. The estimate of L rests on Storey's tail estimator $\widehat{\pi}_0$, whose upward bias is conservative for the FDR but anti-conservative for the FVR (Proposition 2). And applied tetrad analyses correct for multiplicity on the rejection side (Gudergan et al., 2008), which in a validation problem makes acceptance more permissive rather than more careful (Appendix E.2). Extended related work and a method comparison are in Appendix E.3.

2 Method

PCL with known proxies. We observe i.i.d. samples of (T, O, X, M): treatment T, outcome O, observed covariates X, candidate proxies $M = (M_1, \ldots, M_p)$, and an unmeasured confounder U that, with X, would suffice for adjustment (Miao et al., 2018; Tchetgen Tchetgen et al., 2020); we take U scalar, the case in which tetrads are informative (Section 2.1). PCL identifies the effect β of T on O from a negative-control exposure Z and a negative-control outcome W with $O \perp\!\!\!\perp Z \mid (T, U, X)$ and $W \perp\!\!\!\perp (T, Z) \mid (U, X)$; in the linear model, with σ_{jk} the covariance after residualization on X (Appendix A),

$$\beta = \frac{\sigma_{TO}\sigma_{ZW} - \sigma_{TW}\sigma_{ZO}}{\sigma_{TT}\sigma_{ZW} - \sigma_{TW}\sigma_{TZ}} \qquad \text{provided} \qquad \mathcal{D} = \sigma_{TT}\sigma_{ZW} - \sigma_{TW}\sigma_{TZ} = \sigma_{TT}\,\sigma_{ZW\cdot T} \neq 0, \quad (1)$$

where $\sigma_{ZW\cdot T}$ is the partial covariance of Z and W given T. The condition $\mathcal{D} \neq 0$ is the linear form of the relevance (completeness) requirement, and the plug-in estimator $\widehat{\beta}^{\mathrm{PCL}}(Z, W)$ has asymptotic variance of order $\mathcal{D}^{-2}$, so weak relevance, a small $\mathcal{D}$, makes it unstable (Appendix A.5).

2.1 FROM PROXY DISCOVERY TO ONE TEST PER TRIPLET

In our setting the proxies are not known and must be found among the candidates $M_1, \dots, M_p$. We call a measured variable a *disconnected negative control* (DNC) for (T, O) if its only connection to T, to O and to the other candidates runs through U: it is caused by U and it neither causes nor is caused by T, O or another DNC. A DNC is at once a valid negative-control exposure and a valid negative-control outcome, so two DNCs Z and W identify β through equation 1 in either assignment: unlike general PCL proxies, their roles are symmetric (Appendix B.2). This symmetry, and the absence of a causal effect of a DNC on O, have observable implications that motivate the consistency diagnostics of Section 4. Under the DNCT structure, relevance also has a simple characterization. For two DNCs, $\mathcal{D} = \sigma_{ZW}\mathrm{Var}(\varepsilon_T)$ (Appendix A.4), so $\mathcal{D} \neq 0$ exactly when the two are correlated, that is, when both load on U. We therefore admit only candidates that are correlated with one another and with T and O (Sections 3 and 4 give the thresholds); for DNCs this is the relevance condition, and it excludes triplets whose tetrads vanish only because their members carry no signal about U.

DANCE (Kummerfeld et al., 2024) makes the structure testable by working with candidate triplets $S_i = \{A_i, B_i, C_i\} \subset \{M_1, \dots, M_p\}$. In a linear SEM with a single latent confounder U and independent errors, S_i is a *disconnected negative-control triplet* (DNCT) for (T, O) if its three members are DNCs for (T, O) sharing the same U: every trek between a member of S_i and any other variable in $\{A_i, B_i, C_i, T, O\}$ passes through U, and the edge $T \to O$ is allowed (Figure 1A; Appendix B.1). Every ordered pair of a DNCT identifies β, so a DNCT supplies six proxy assignments with one population estimand. The candidates are all $\binom{p}{3}$ triplets of a pool. We write $H_i = 0$ if S_i is a DNCT and $H_i = 1$ otherwise, and throughout call $H_i = 0$ candidates *valid* (structurally valid), $H_i = 1$ candidates *invalid*, a candidate *tetrad-compatible* when its population tetrads vanish (Appendix B.5), and *accepted* when a statistical rule selects it, whatever its truth.

Vanishing tetrads. Under a DNCT the variables A_i, B_i, C_i are connected to T and O only through the scalar latent confounder U, apart from the allowed edge $T \to O$ (Figure 1A). Consequently, within each of the two sets $\{A_i, B_i, C_i, T\}$ and $\{A_i, B_i, C_i, O\}$ the cross-covariances have one-factor form $\sigma_{jk} = \lambda_j \lambda_k \mathrm{Var}(U)$ (Appendix B.2), and every *tetrad*, the 2×2 minor between two disjoint pairs, vanishes; for example $T_{AB|CT} = \sigma_{AC}\sigma_{BT} - \sigma_{AT}\sigma_{BC} = 0$. Each four-variable set has three such tetrads, giving the six constraints used by DANCE: three *T-anchored* and three *O-anchored* tetrads (Figure 1B; Appendix B.3 gives a four-variable primer and Appendix B.4 writes all six out). The remaining tetrads of the five variables either involve σ_{TO}, and therefore need not vanish in the presence of the causal edge $T \to O$, or are implied by these six when the candidates are correlated (Appendix B.4). Within each family one tetrad is an exact linear combination of the other two, so the six constraints are equivalent to $t(\Sigma_i) = (T_{AB|CT}, T_{AC|BT}, T_{AB|CO}, T_{AC|BO})^\top = \mathbf{0}$, and we test the composite null $\mathcal{H}_{0i}^{\mathrm{tet}} : t(\Sigma_i) = \mathbf{0}$. A DNCT necessarily satisfies $\mathcal{H}_{0i}^{\mathrm{tet}}$, but the converse requires *tetrad faithfulness*: an invalid triplet must not satisfy the same rank-one covariance constraints by parameter cancellation. Thus the test identifies tetrad-compatible triplets, and identifies DNCTs only under faithfulness (Appendix B.5). For example, a second latent factor whose loadings on A_i, B_i, C_i are proportional to those of U is invisible to the tetrads, even if it also affects O and invalidates PCL. More generally, non-proportional loadings from multiple latent confounders produce higher-rank covariance blocks and generically violate the tetrad constraints; such settings require larger proxy sets and rank tests (Appendix B.5; Xie et al., 2024).

One joint test per triplet (Figure 1B). DANCE tests the six tetrad constraints of a candidate triplet separately. For each tetrad it uses Wishart's standardized statistic and obtains one p-value; a triplet is retained only if all six exceed the threshold α (default $\alpha = 1/N$; Wishart, 1928; Appendix B.7), a rule we call *DANCE all-pass*. We instead treat the triplet as the statistical unit and test the single composite null $\mathcal{H}_{0i}^{\mathrm{tet}}$. Let $\hat{t}_i = t(\widehat{\Sigma}_i)$ and let $\widehat{\Sigma}_{t,i}$ estimate the asymptotic covariance $\Sigma_{t,i}$ of $\sqrt{N}\,\hat{t}_i$ (Appendix B.6). The simultaneous tetrad test of Bollen (1990) is

$$T_i = N\,\hat{t}_i^\top \widehat{\Sigma}_{t,i}^{+}\,\hat{t}_i, \qquad P_i = 1 - F_{\chi^2_{q_i}}(T_i), \qquad q_i = \mathrm{rank}(\Sigma_{t,i}), \tag{2}$$

for which $T_i \xrightarrow{d} \chi^2_{q_i}$ under $\mathcal{H}_{0i}^{\mathrm{tet}}$ (Figures 2B and 8B); $q_i = 4$ in all reported analyses. Thus each candidate triplet receives one asymptotically calibrated p-value for the joint tetrad constraints. Under a fixed alternative the statistic is approximately noncentral $\chi^2_{q_i}(\nu_i)$ with $\nu_i = N\,t(\Sigma_i)^\top \Sigma_{t,i}^{+} t(\Sigma_i)$, so sample size and distance from the tetrad null enter detectability as one number that grows linearly in N

(Appendix B.6). Accounting for $\Sigma_{t,i}$ is essential because the four tetrads share sample covariances and are statistically dependent even though they are algebraically independent (Appendix B.6, Table 1). The candidate-level p-value is also the appropriate input to the two-group model of Section 2.2: under tetrad faithfulness its π_0 is the proportion of valid triplets and $L(P_i)$ a candidate-level validity score; without faithfulness these quantities instead refer to tetrad compatibility. Applying the mixture model to DANCE's six component p-values, as the comparator *Wishart-component* of Section 3 does, would make π_0 a proportion of vanishing tetrads rather than of candidate triplets and would ignore their within-triplet dependence (Appendix B.7). Dependence *between* overlapping triplets remains and is studied separately in Section 3.2.

2.2 Local validity and accepted-set FVR

Two-group model. Let $\pi_0 = \mathbb{P}(H_i = 0)$ be the proportion of valid triplets among the m candidates. A valid triplet is tetrad-compatible and the joint test is calibrated (Figure 8B), so $P_i \mid H_i = 0$ is asymptotically Uniform$(0, 1)$. An invalid triplet with noncentrality $\nu_i > 0$ has a p-value density f_1 concentrated near zero and decreasing toward one (Figure 2B; Appendix C.1), so across candidates the p-values follow the two-group mixture (Storey, 2002; Storey & Tibshirani, 2003)

$$f_P(p) = \pi_0 \cdot 1 + (1 - \pi_0) f_1(p), \tag{3}$$

and every quantity below is a function of (π_0, f_1). For an accepted set $\mathcal{A}$, $\mathrm{FVR}(\mathcal{A}) = \mathbb{P}(H_i = 1 \mid i \in \mathcal{A})$ is the probability that an accepted triplet is invalid; we refer to it as the *contamination* of $\mathcal{A}$. Retention is the valid-triplet coverage $C_{\mathrm{val}}(\mathcal{A}) = \mathbb{P}(i \in \mathcal{A} \mid H_i = 0)$.

The rule "accept $P_i > \gamma$" accepts a valid triplet with probability $1 - \gamma$ and an invalid one with probability $\bar{F}_1(\gamma) = \int_\gamma^1 f_1$, so

$$\mathrm{FVR}(\gamma) = \mathbb{P}(H_i = 1 \mid P_i > \gamma) = \frac{(1 - \pi_0)\bar{F}_1(\gamma)}{\pi_0(1 - \gamma) + (1 - \pi_0)\bar{F}_1(\gamma)} = \frac{\int_\gamma^1 (1 - \pi_0) f_1(p)\, dp}{\int_\gamma^1 f_P(p)\, dp}, \tag{4}$$

the accepted invalid triplets as a share of all accepted triplets. $\mathrm{FVR}(\gamma)$ decreases in π_0 and increases in $\bar{F}_1(\gamma)$: the same cutoff gives a low FVR when valid triplets are common and violations detectable, a high one when valid triplets are rare or violations weak (Appendix C.2).

Local validity. Instead of conditioning on the event $\{P > \gamma\}$, condition on the p-value itself. Bayes' rule on a shrinking interval $[p, p + dp]$ (Appendix C.3) gives the *local validity*

$$L(p) := \mathbb{P}(H_i = 0 \mid P_i = p) = \frac{\pi_0}{f_P(p)}, \qquad \text{equivalently} \qquad 1 - L(p) = \frac{(1 - \pi_0) f_1(p)}{f_P(p)}, \tag{5}$$

the posterior probability that a triplet with p-value p is valid. The second form is the pointwise version of equation 4: the share of the density at p that comes from invalid triplets. The two are linked for any acceptance region $\mathcal{R}$ by $\mathrm{FVR}(\mathcal{R}) = \mathbb{E}[1 - L(P_i) \mid P_i \in \mathcal{R}]$ (tower property; Appendix C.4), so the FVR of a rule is the average of $1 - L$ over the p-values it accepts. $L(p)$ is close to one where $f_P(p)$ is close to its floor π_0, that is, where the invalid density has essentially vanished. Numerically L is the local false discovery rate (Efron et al., 2001; Efron, 2010). The oracle rule accepts $\mathcal{A}^*_\tau = \{i : L(P_i) \geq \tau\}$.

Proposition 1 (Oracle accepted-set FVR). *If* $\mathbb{P}(i \in \mathcal{A}^*_\tau) > 0$, *then* $\mathrm{FVR}(\mathcal{A}^*_\tau) = 1 - \mathbb{E}[L(P_i) \mid L(P_i) \geq \tau] \leq 1 - \tau$.

This is the identity above on the region $\{L \geq \tau\}$, where $1 - L \leq 1 - \tau$. The statement is marginal and uses no independence between triplets; under an i.i.d. two-group model the realized fraction I/R of invalid triplets among the R accepted also satisfies $\mathbb{E}[I/R \mid R > 0] \leq 1 - \tau$, which is what the simulations report over nonempty accepted sets, an empty set counting as an abstention rather than as zero contamination (Appendix C.4).

Why the oracle validity threshold transports and a raw-p cutoff does not. Since L is nondecreasing in p whenever f_1 is nonincreasing, as it is for noncentral χ^2 alternatives (Appendix C.8), the rule $L(P_i) \geq \tau$ accepts exactly the p-values above the cutoff c_τ at which $L(c_\tau) = \tau$, that is, $f_1(c_\tau) = \pi_0(1 - \tau)/\{\tau(1 - \pi_0)\}$. The two rules differ in what is held fixed when applied to a

new dataset. A raw-p rule holds γ fixed, and by equation 4 its FVR then changes with (π_0, f_1), so it has to be re-tuned for every dataset. The posterior rule holds the target $1-\tau$ fixed and lets the mixture set the cutoff: c_τ moves toward one when π_0 is small or f_1 is flat, and the rule accepts fewer triplets, possibly none. For the oracle rule, the price of a fixed target is paid in coverage rather than contamination (Figure 2E; worked example in Appendix C.5).

Plug-in estimation. In practice (π_0, f_P) are estimated from the m candidate p-values with `qvalue` (Storey & Tibshirani, 2003; Storey et al., 2023) (Appendix C.6): $\widehat{\pi}_0$ is the spline-smoothed form of Storey's tail estimator, $\widehat{f}_P$ a kernel density estimate, $\widehat{L}(p) = \min\{1, \widehat{\pi}_0/\widehat{f}_P(p)\}$ made nondecreasing in p, and $\widehat{\mathcal{A}}_\tau = \{i : \widehat{L}(P_i) \geq \tau\} = \{i : P_i \geq \hat{c}_\tau\}$, with headline threshold $\tau = 0.90$. The oracle bound does not transfer automatically: if $\sup_p |\widehat{L}(p) - L(p)| \leq \varepsilon$ then $\mathrm{FVR}(\widehat{\mathcal{A}}_\tau) \leq 1 - \tau + \varepsilon$ (Appendix C.7), so what matters is the size and direction of the estimation error. The underlying tail estimator $\widehat{\pi}_0(\lambda) = \#\{P_i > \lambda\}/\{m(1-\lambda)\}$ reads the floor from the right tail, where valid triplets contribute a fraction $\pi_0(1-\lambda)$ of all p-values; if invalid p-values land there too, $\mathbb{E}\,\widehat{\pi}_0(\lambda) = \pi_0 + (1-\pi_0)\bar{F}_1(\lambda)/(1-\lambda) > \pi_0$. For the FDR this upward bias is conservative; for the FVR it is not, because a larger $\widehat{\pi}_0$ raises every $\widehat{L}$ and enlarges $\widehat{\mathcal{A}}_\tau$: upward contamination of the estimated null component, harmless for rejection-side FDR control, reverses direction when the nulls are what we accept.

Proposition 2 (Acceptance-side purity)**.** *Let $\theta = \operatorname{ess\,inf}_p f_1(p)$ and $\pi_0^\star = \pi_0 + (1-\pi_0)\theta$. (i) $\pi_0^\star = \operatorname{ess\,inf}_p f_P(p)$ is the largest uniform weight compatible with f_P: with f_1 unrestricted, f_P identifies $\pi_0^\star$ but not π_0, and $\mathbb{E}\,\widehat{\pi}_0(\lambda) \geq \pi_0^\star$ for every $\lambda \in (0,1)$. (ii) If the rule accepting $L^\star(p) = \pi_0^\star/f_P(p) \geq \tau$ accepts with positive probability, then $1 - \pi_0/\pi_0^\star \leq \mathrm{FVR} \leq 1 - \tau\,\pi_0/\pi_0^\star$.*

Proofs, and the value of θ under noncentral χ^2 alternatives, are in Appendix C.8. The content of the proposition is that some invalid mass is observationally indistinguishable from null mass: θ is the part of the invalid density that is flat over the whole interval and therefore indistinguishable from the valid floor; the two-group models (π_0, f_1) and $(\pi_0^\star, (f_1 - \theta)/(1-\theta))$ produce the identical f_P, so no procedure that sees only the p-values, whether it thresholds $\widehat{L}$, a q-value or a raw p-value, can tell them apart, and every rule inherits the same bias. This is the impurity of Genovese & Wasserman (2004), anti-conservative on the acceptance side. Its size is set by the least detectable invalid triplets: ν_i grows linearly in N, so $\theta \to 0$ for any detectable violation, whereas tetrad-faithfulness failures ($\nu_i = 0$) contribute to θ at every N (Appendix B.5). In the very-small regime at $N = 2000$ ($\theta \approx 0.36$) part (ii) places the FVR at $\tau = 0.9$ in $[0.60, 0.64]$, which is what the plug-in rule reaches there (Figure 2D); where θ is negligible the mean realized FVR is at or below the target (Appendix C.8 works through the regimes).

2.3 From false validation to population structural bias

Accepted triplets are passed to PCL. Write the members of S_i as Z_{i1}, Z_{i2}, Z_{i3}; each ordered pair (Z_{ia}, Z_{ib}) is a candidate (Z, W) assignment, and since $\mathcal{D}$ in equation 1 is symmetric in (Z, W), relevance is a property of the three unordered pairs. For a valid triplet relevance reduces to the proxies being correlated, which every candidate pool requires before testing (Section 2.1); for an invalid triplet $\mathcal{D}$ is unconstrained. When all three pairs satisfy $\mathcal{D} \neq 0$ we average the six ordered-pair population estimands, $\bar{\beta}_i = \frac{1}{6}\sum_{a \neq b} \beta^{\mathrm{PCL}}(Z_{ia}, Z_{ib})$, where $\beta^{\mathrm{PCL}}(Z, W)$ is the right-hand side of equation 1, and write $b_i = \bar{\beta}_i - \beta$. The population structural bias b_i is zero for every valid triplet, because each ordered pair identifies β (Appendix A.3). An invalid triplet may induce large, small or zero bias; the tetrad test sees the tetrads, not b_i (finite-sample details in Appendix C.10).

For a realized accepted set $\mathcal{A} \neq \emptyset$, let $I(\mathcal{A}) = \sum_{i \in \mathcal{A}} \mathbf{1}\{H_i = 1\}$ be the number of accepted invalid triplets and $\widehat{\mathrm{FVR}}(\mathcal{A}) = I(\mathcal{A})/|\mathcal{A}|$ the realized contamination. When $I(\mathcal{A}) > 0$, let $\mathrm{Sev}(\mathcal{A}) = I(\mathcal{A})^{-1} \sum_{i \in \mathcal{A},\, H_i = 1} |b_i|$ be their mean absolute bias, the severity of the violations among the accepted triplets. Since $b_i = 0$ for every valid triplet, the mean absolute bias over the whole accepted set factorizes:

$$\frac{1}{|\mathcal{A}|}\sum_{i \in \mathcal{A}} |b_i| = \widehat{\mathrm{FVR}}(\mathcal{A})\ \mathrm{Sev}(\mathcal{A}), \tag{6}$$

with both sides zero when $I(\mathcal{A}) = 0$. The identity is exact and holds for every accepted set, whatever rule produced it. It separates two questions that are otherwise conflated: how often invalid triplets are

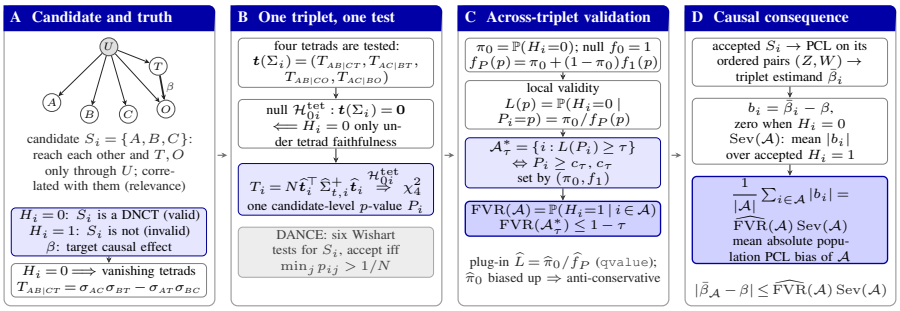


Figure 1: **Method at a glance. A:** a candidate triplet of disconnected negative controls; $H_i = 0$ valid, $H_i = 1$ invalid. **B:** validity implies four independent vanishing tetrads (converse under tetrad faithfulness); Bollen's joint Wald test turns them into one candidate-level p-value. **C:** a two-group model across candidates converts P_i into the local validity $L(p)$; the oracle rule $L(P_i) \geq \tau$ has FVR at most $1-\tau$. **D:** accepted candidates enter PCL; their mean absolute population bias equals empirical FVR times severity. In symbols: $H_i = 0 \Rightarrow t(\Sigma_i) = \mathbf{0}$; $t \to T_i \to P_i$; $P_i \to L(P_i) \to \mathcal{A}_\tau$; $\mathcal{A}_\tau \to$ PCL with bias $=$ FVR $\times$ severity.

accepted, which depends on the screening rule and on how well valid and invalid p-values separate, and how harmful the accepted invalid triplets are, which depends on the data-generating structure and is beyond the reach of any screening rule; Appendix D reports the product on a grid that varies the two separately (Figure 7). Two consequences follow. First, the identity bounds the average of $|b_i|$, whereas the pooled estimand $\bar{\beta}_{\mathcal{A}} = |\mathcal{A}|^{-1}\sum_{i\in\mathcal{A}} \bar{\beta}_i$ averages the signed biases, which may partly cancel. By the triangle inequality, $|\bar{\beta}_{\mathcal{A}} - \beta| \leq \widehat{\mathrm{FVR}}(\mathcal{A})\ \mathrm{Sev}(\mathcal{A})$: contamination times severity bounds the bias of the pooled estimate, with equality when all accepted invalid triplets are biased in the same direction. Because cancellation depends on unknown signs and would flatter any screening rule, our benchmark is built as that worst case, with λ chosen so that every invalid bias has the same sign (Appendix D.2). Second, the identity is about population estimands; the estimate $\widehat{\bar{\beta}}_{\mathcal{A}}$ differs from $\bar{\beta}_{\mathcal{A}}$ by sampling error and, because the same data select $\mathcal{A}$ and estimate β, by post-selection error. The simulations measure both parts (Section 3.2; Appendix D.4). Proofs, the signed version and the finite-sample decomposition are in Appendix C.10.

3 Simulations

The simulations know each H_i and β, so each stage can be checked separately: (I) calibration and informativeness of the p-values; (II) whether the plug-in rule attains its target and whether one threshold transports; (III) the effect of dependence on the mixture fit; (IV) the effect of accepted invalid triplets on PCL. Designs, full results and panel-by-panel computations are in Appendix D. In replicate r a rule accepts R_r triplets, I_r of them invalid, and we report $\widehat{\mathrm{FVR}}_r = I_r/R_r$ over replicates with $R_r > 0$. A replicate with $R_r = 0$ is an *abstention*, never zero contamination (Appendix C.4), and a failed `qvalue` fit is recorded separately. Valid-triplet coverage is the fraction of valid triplets accepted, averaged over all replicates with abstentions counted as zero.

V-factor benchmark (Figure 2A). Data follow a linear SEM with two independent latent factors. U is the confounder: every gene, the treatment and the outcome load on it. V is a second factor that loads on some genes and on the outcome, but not on the treatment:

$$n_j = u_j U + \varepsilon_j, \qquad g_i = u_i U + v_i V + \varepsilon_i, \qquad T = aU + \varepsilon_T, \qquad O = \beta T + cU + \lambda V + \varepsilon_O. \quad (7)$$

Each coefficient has one role. u_i is proxy strength, how well gene i measures U; it is evenly spaced on $[0.8, 1.5]$ and the same in every cell. a and c are the confounding paths $U \to T$ and $U \to O$ (both 1.15), and $\beta = 0.5$ is the effect to be recovered. v_i is the loading on V, and it is what makes a gene invalid: genes with $v_i = 0$ (the valid panel, n_j) reach T and O only through U, so any three of them form a DNCT with all tetrads zero and PCL bias $b_i = 0$; genes with $v_i \neq 0$ (the violation panel, g_i) also reach O through V, which breaks the tetrads and, through the path $g_i \leftarrow V \to O$, biases PCL when they are used as proxies. We write $v_i = r_i u_i$, so that the ratio r_i measures how strongly a gene reflects V relative to U whatever its proxy strength. λ is the $V \to O$ loading: it sets how harmful an

invalid triplet is and, because it also enters the O-anchored tetrads, affects detectability too. Three things vary across cells, nothing else. The *violation regime* is the range of the ratios r_i in the violation panel: near null $(10^{-7}, 5 \times 10^{-6})$, very small $(0.05, 0.15)$, small $(0.15, 0.4)$, moderate $(0.4, 0.9)$, large $(0.9, 1.8)$. Since every tetrad of an invalid triplet is proportional to a difference of ratios, the regime fixes how far invalid triplets sit from the tetrad null, from indistinguishable at any N (near null, where the PCL bias is of the same 10^{-10} order and the regime needs no screening) to detected at every N studied (moderate, large), with very small and small in between (Appendix D.2 gives the noncentralities); within a regime the ratios are evenly spaced, so no two genes share a ratio and no triplet sits on the rank-one blind spot. The sample size N runs from 500 to 50,000. And π_0 is set by the panel sizes, because candidates are pure triplets, all three genes from one panel: at $\pi_0 = 0.191$, 14 valid and 22 violation genes give 364 valid and 1,540 invalid triplets, $m = 1{,}904$. λ is set from the regime, $\lambda = 1.5\, c \max_i r_i$, so that every invalid triplet's bias has the same sign and cancellation cannot flatter the screening (Section 2.3). All-violation triplets are the hard case for contamination: in mixed panels, triplets with one or two violation genes carry larger tetrad signal and are accepted at rates $\leq 10^{-4}$ (Appendix D.2).

3.1 Claims I and II: calibration and acceptance-side validation (Figure 2)

Calibration and separation (B). For valid triplets the joint test is slightly conservative at $N = 500$ (mean size 0.041 at level 0.05) and within 0.005 of nominal for $N \geq 2000$, and the null QQ curve follows the diagonal (Figure 8B). In the near-null regime $f_1 \approx f_0$ and the p-values carry no class information.

Recovery of π_0 (C). Wherever f_1 separates from f_0 the mixture is centred on the truth: the mean error of $\widehat{\pi}_0$ is within 0.02 at every N in the moderate and large regimes and from $N = 2000$ in the small regime. In the very-small regime the error falls from 0.55 at $N = 500$ to 0.02 at 5×10^4: the upward bias of Proposition 2(i), shrinking as ν_i grows with N. In the near-null regime $\widehat{\pi}_0 \approx 0.88$ whatever the truth, the correct answer to how many p-values look null when $f_1 \approx f_0$; the Wishart-component mixture overestimates π_0 in every cell (Figure 8C).

Does $\tau = 0.90$ deliver FVR ≤ 0.10? (D). Panel D shows $\widehat{\mathrm{FVR}} - 0.10$ for the plug-in and oracle rules. The mean realized FVR of the plug-in rule is at or below the target in the moderate and large regimes. In the small regime its FVR falls from 0.31 at $N = 1000$ to 0.036 at $N = 10^4$, approaching the oracle value 0.030; where it misses, invalid p-values keep mass near one, the impurity of Proposition 2, and at $N = 1000$ a plug-in score of 0.99 corresponds to only 69% valid triplets. In the very-small regime it stays above 0.4 at every N (Appendix D.4). DANCE all-pass is heavily contaminated in the same cells (Figure 8D). The price of meeting the target is coverage: the plug-in rule retains 55–60% of valid triplets in the moderate and large regimes and, in the small regime, abstains in 44–52% of replicates for $N \leq 5000$ (Figure 8E, G).

Why not a fixed raw-p cutoff? (E). In a paired design that holds f_0 and f_1 fixed and varies only π_0 and N (Appendix D), the cutoff $c^*_{0.10}$ at which $P_i > c$ attains FVR 0.10 moves by more than two orders of magnitude with N, shifts with π_0, and is unattainable in the weakest regimes; a cutoff $c = 0.075$ tuned on the identifiable cells keeps coverage near 0.93 on held-out replicates but yields contamination from about 0 to 0.7 across cells.

Downstream causal-effect bias, preview (F). Without screening the PCL estimate is biased by -0.05 to -0.66 from the very-small to the large regime ($\beta = 0.5$); after screening the bias is near zero for $N \geq 5000$ in every regime.

3.2 Claims III and IV: dependence and downstream causal bias (Figure 7)

Dependence reduces the information available to the mixture (A, B). Triplets that share genes have dependent p-values. Since changing the overlap in the V-factor design would also change the tests themselves, we vary dependence alone, at the p-value level, through a block-equicorrelated Gaussian copula that leaves every marginal unchanged, and summarize its effect by $m_{\mathrm{eff}} = m\mathrm{Var}(\widehat{\pi}_0 \mid \rho = 0)/\mathrm{Var}(\widehat{\pi}_0 \mid \rho)$, the number of independent tests that would estimate π_0 as precisely as the m dependent ones. For every m, m_{eff}/m falls from about 0.7 at $\rho = 0.1$ to 0.12 at $\rho = 0.9$ (A).

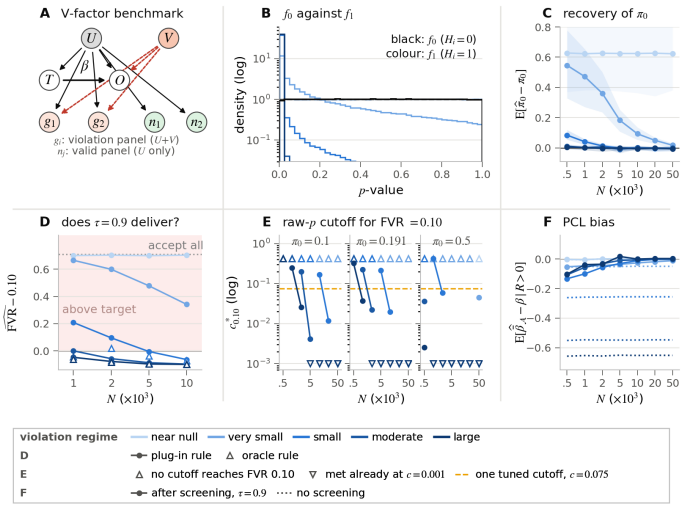


Figure 2: **End-to-end behaviour of the validation pipeline** (V-factor benchmark, $\beta = 0.5$, 250 replicates per cell; colours are the five violation regimes). **A:** design. **B:** p-value densities of valid (f_0, black) and invalid (f_1, one curve per regime) triplets at $\pi_0 = 0.191$, $N = 5000$. **C:** mean signed error of $\widehat{\pi}_0$ over $\pi_0 \in \{0, 0.1, 0.191, 0.308, 0.407, 0.5\}$ (band: range). **D:** $\widehat{\mathrm{FVR}} - 0.10$ at $\tau = 0.9$ for the plug-in rule (filled) and the oracle rule (open triangles; true π_0, f_1 from independent data), $\pi_0 = 0.191$, $m = 1904$. **E:** raw-p cutoff $c^*_{0.10}$ attaining FVR 0.10 in the paired-list design ($m = 2000$), one facet per π_0. **F:** signed bias of $\widehat{\widehat{\beta}}_{\mathcal{A}}$ over nonempty accepted sets, with and without screening. Comparators, coverage, abstention, null calibration: Figure 8.

Dependence leaves $\widehat{\pi}_0$ centred and the mean FVR near target but makes both more variable: the probability that a nonempty accepted set misses the 10% target falls to 0.06 at $m = 3000$ under independence but stays near 0.4–0.5 at every m for $\rho \geq 0.6$ (B). Inside the SEM, the 7140 overlapping triplets of a 36-gene panel estimate π_0 only as precisely as 220 gene-disjoint ones (Appendix D.4): the nominal candidate count overstates the information available to the mixture (Appendix C.9).

Downstream bias in the estimated effect (C, D). We rerun the pipeline on a grid that separates detectability, controlled mainly by the scale δ of the V-loadings, from the severity of the accepted violations, controlled mainly by the $V \to O$ loading λ (not orthogonal, since λ also enters the O-anchored tetrads; Appendix D.2). Panel C reports the mean of $\widehat{\mathrm{FVR}}(\mathcal{A})$ $\mathrm{Sev}(\mathcal{A})$, the accepted set's mean absolute structural bias equation 6. It is large only where invalid triplets are both accepted and harmful: at most 0.015 for $\delta \geq 1$, against 0.29 at $(\delta, \lambda) = (0.25, 3)$. Panel D reports the RMSE of $\widehat{\widehat{\beta}}_{\mathcal{A}}$, from 0.02 to 0.32, which exceeds the structural term in every cell because it also contains sampling error. The systematic part of the estimation-and-selection error is small: on this grid the mean signed error of $\widehat{\widehat{\beta}}_{\mathcal{A}}$ is within 0.013 of the signed structural term in every cell, and where that term exceeds 0.02 the RMSE is 1.1 to 1.7 times larger; the pure-triplet grid shows the same pattern with larger gaps at $N \leq 2000$ (Appendix D.4). Screening can make the bias from accepted invalid triplets negligible, but not the ordinary estimation error. Severity sweeps, sign-cancellation checks and the frozen-selection verification of equation 6 are in Appendix D.4.

4 TCGA–BRCA APPLICATION

We use TCGA–BRCA RNA-seq data from $N = 507$ tumours as an illustrative high-dimensional application of the discovery pipeline, not as a definitive causal analysis of gene expression on stage: AJCC stage is the outcome O and, in turn, each candidate gene is the treatment T. Because relevance requires correlated proxies (Section 2.1), candidate pools are correlation-coherent clusters in which every pair of genes, the treatment included, has $|\widehat{r}| \geq 0.35$. Of 611 panel genes, 203 have pools of at least four members (median 7), giving $m = 5{,}710{,}820$ candidate triplets, each tested with $q = 4$ (Appendix F).

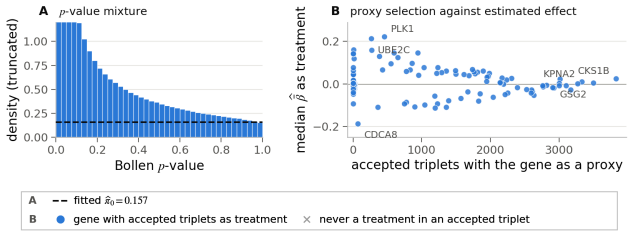


Figure 3: **TCGA–BRCA application. A:** density of the 5,710,820 candidate-level Bollen p-values (axis truncated at 1.2); dashed: fitted $\widehat{\pi}_0 = 0.157$. **B:** for each gene, the number of accepted triplets ($\widehat{L} \geq 0.9$; 45,270 in all) in which it appears as a proxy, against the median PCL estimate over the accepted triplets in which it is the treatment; crosses: genes never the treatment of an accepted triplet; the most selected proxies and the largest estimates are labelled.

The p-value distribution (Figure 3A) has a large concentration near zero and an approximately flat right tail whose height matches the fitted $\widehat{\pi}_0 = 0.157$, as the two-group model predicts. At $\tau = 0.9$ the rule accepts 45,270 triplets, 0.8% of the candidates; DANCE's rule, smallest Wishart p-value above $1/N$, accepts 3,198,500, or 56%. Four caveats govern these numbers. Structural validity is unobserved, so $\widehat{\pi}_0$ and the accepted set refer to tetrad-compatibility under faithfulness. At $N = 507$ weak violations separate poorly from the null, the regime in which $\widehat{\pi}_0$ is overestimated and the accepted set is more contaminated than $1 - \tau$ (Proposition 2; Figure 2D), so the accepted set is a ranked candidate list, not a certified one. And the 5.7 million triplets are built from 611 genes, so the information behind the mixture fit is far smaller than m (Section 3.2). And the analysis uses the normal-theory tetrad covariance on log-scale expression and numerically coded stage; null calibration in this non-Gaussian setting is not guaranteed, so the application should be read as illustrative, and a sandwich or permutation calibration is a natural extension (Appendix B.9).

Two consistency checks follow from the DNC definition (Section 2.1). A DNC has no causal effect on O, so a gene often accepted as a proxy should have a PCL estimate near zero when it is itself the treatment: the seven most selected proxies, each in 3,000–3,800 accepted triplets, have median treatment estimates within ± 0.03 of zero, whereas the largest estimates (0.22 for PLK1 and UBE2C, -0.19 for CDCA8) belong to genes selected fewer than 500 times (Figure 3B; Appendix F.4). Second, two DNCs identify the same estimand in either assignment, so the six ordered-pair estimates of an accepted triplet should agree; their within-triplet standard deviation has median 0.017 among accepted triplets against 0.062 among rejected ones, a contrast that persists after stratification by denominator strength (Appendix F.5, Figure 9). Both are checks of coherence, not validation.

5 Discussion

We have treated proxy discovery as an acceptance-side multiple-testing problem inside causal estimation, with one tetrad p-value per candidate, a validity score whose oracle rule has FVR $\leq 1-\tau$, and an identity linking the accepted set's bias to FVR times severity.

Limitations and outlook. The guarantees are conditional on what the test can see. Tetrad faithfulness is untestable, vanishing tetrads need an effectively one-dimensional latent structure (Appendix B.5), and several confounders call for rank constraints (Xie et al., 2024); the acceptance-side layer itself is agnostic to tetrads: any calibrated candidate-level structural test whose p-values are stochastically smaller under violations can replace Bollen's statistic (Appendix E.4), so nonlinear settings can substitute the generalized tetrad statistic (Ying et al., 2025). The tetrad test targets structural compatibility, whereas relevance is imposed separately through the correlation requirement (Appendix A.4). The plug-in rule's contamination is quantified, not guaranteed: it can exceed $1 - \tau$ where invalid p-values keep mass near one (Proposition 2), the situation at small N and weak violations and hence at the TCGA sample size, where the accepted set is a ranked candidate list. Finally, selection and estimation share the data, so the reported PCL errors carry post-selection error; selective inference for the accepted-set estimand (Fithian et al., 2017; Lee et al., 2016) is the natural next step.

REPRODUCIBILITY STATEMENT

Every simulation number and figure comes from one Python package (`vfactor_sim`) with local false discovery rates from the R package `qvalue` through a fixed interface; parameters are deterministic functions of the configuration and seeds are derived from configuration, regime, sample size and replicate index, so each cell reproduces draw for draw. Appendix D specifies every experiment, Appendix D.4 lists the numbers quoted in the text, Appendix D.5 states how each panel is computed, Appendix B.6 gives the implementation constants of the joint test (ridge 10^{-7}, rank tolerance 10^{-10}, pseudo-inverse), Appendix C.6 the `qvalue` settings, and Appendix F the TCGA–BRCA protocol, whose released code reproduces the reported statistics to 2×10^{-11}. Proofs are in Appendices C and C.10. Code, the pinned TCGA gene list and per-cell result tables are released with the paper.

AI USAGE DISCLOSURE

In this work, we used generative AI tools for generating parts of the simulation and analysis code, which the authors verified and tested for correctness. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims. Additionally, we used generative AI tools for draft parts of a research paper, summarize or analyse existing literature, edit a research paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Kenneth A. Bollen. Outlier screening and a distribution-free test for vanishing tetrads. *Sociological Methods & Research*, 19(1):80–92, 1990.

Kenneth A. Bollen and Kwok-fai Ting. Confirmatory tetrad analysis. In *Sociological Methodology*, volume 23, pp. 147–175. Wiley, 1993.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024. First circulated 2020.

Bradley Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

Bradley Efron, Robert Tibshirani, John D. Storey, and Virginia Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001. doi: 10.1198/016214501753382129.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2017.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B*, 64(3):499–517, 2002.

Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32(3):1035–1061, 2004.

Siegfried P. Gudergan, Christian M. Ringle, Sven Wende, and Anja Will. Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research*, 61(12):1238–1249, 2008.

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.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

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.

Sanat K. Sarkar. False discovery and false nondiscovery rates in single-step multiple testing procedures. *The Annals of Statistics*, 34(1):394–415, 2006.

Albert Satorra and Peter M. Bentler. Corrections to test statistics and standard errors in covariance structure analysis. *Latent Variables Analysis: Applications for Developmental Research*, pp. 399–419, 1994.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

Charles Spearman. General intelligence, objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64(3):479–498, 2002.

John D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

John D. Storey, Andrew J. Bass, Alan Dabney, and David Robinson. *qvalue: Q-value estimation for false discovery rate control*, 2023. URL https://bioconductor.org/packages/qvalue. R package version 2.34.0.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

John Wishart. Sampling errors in the theory of two factors. *British Journal of Psychology*, 19: 180–187, 1928.

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*. PMLR, 2024.

Naiwen Ying, Ping Zhang, Shanshan Luo, and Wang Miao. A generalized tetrad constraint for testing conditional independence given a latent variable. *arXiv preprint arXiv:2504.14173*, 2025.

Supplementary Material

Appendices to "Acceptance-Side Multiple Testing for Proxy Discovery"

These appendices contain the material referenced from the main text: Appendix A (proximal causal learning with known proxies), Appendix B (DNCT structure, tetrad faithfulness and the joint tetrad test), Appendix C (acceptance-side validation: derivations, proofs and worked examples), Appendix D (simulation designs, results and figure guide), Appendix E (terminology, fixed rules, method comparison and modularity) and Appendix F (TCGA–BRCA protocol and diagnostics). Equation, figure and table numbers continue from the main text; section numbers are lettered.

A Proximal causal learning with known proxies

A.1 Setting

We observe i.i.d. copies of (T, O, Z, W, X): the treatment, the outcome, a treatment-inducing proxy Z (negative-control exposure), an outcome-inducing proxy W (negative-control outcome), and baseline covariates X. A confounder U is unobserved. The target is the average treatment effect $\mathbb{E}[O_1 - O_0]$; in the linear model of Appendix A.3 it equals the structural coefficient β. Conditioning on (Z, W) does not remove the confounding: they are noisy proxies for U, not U itself, and the residual variation in U that remains after conditioning still enters both T and O. PCL removes it instead by exploiting how Z and W relate to U.

A.2 Assumptions and identification

Write $O_{t,z}$ and $W_{t,z}$ for the potential outcomes under $\mathrm{do}(T = t, Z = z)$. The three structural assumptions of Miao et al. (2018) and Tchetgen Tchetgen et al. (2020) are the following.

- *Latent exchangeability:* $(O_{t,z}, W_{t,z}) \perp\!\!\!\perp (T, Z) \mid (U, X)$ for all (t, z). If U and X were known, the realized (T, Z) would carry no further information about the potential outcomes.
- *Negative-control exposure:* $O_{t,z} = O_t$ for all (t, z), so intervening on Z with T fixed does not move the outcome; its observed-data implication is $O \perp\!\!\!\perp Z \mid (T, U, X)$. Z may be caused by U and may even affect T, but it provides no route to O other than through (T, U, X).
- *Negative-control outcome:* $W_{t,z} = W$ for all (t, z), so neither T nor Z causes W; its observed-data implication is $W \perp\!\!\!\perp (T, Z) \mid (U, X)$.

Identification proceeds through an outcome confounding bridge h solving

$$\mathbb{E}[O \mid T = t, Z = z, X = x] = \mathbb{E}[h(t, x, W) \mid T = t, Z = z, X = x] \qquad \text{for all } (t, z, x). \tag{8}$$

Two completeness conditions are used. Latent completeness, $\mathbb{E}[v(U) \mid Z, T, X] = 0 \Rightarrow v(U) = 0$ almost surely for every square-integrable v, says that Z carries enough variation in U to distinguish functions of U; in the categorical case it requires $\min(d_Z, d_W) \geq d_U$. A completeness of Z given (W, T, X) guarantees that a bridge exists. Under these conditions any solution h of equation 8 satisfies $\mathbb{E}[O_t] = \mathbb{E}[h(t, X, W)]$, hence

$$\mathrm{ATE} = \mathbb{E}[O_1 - O_0] = \mathbb{E}\big[h(1, X, W) - h(0, X, W)\big]. \tag{9}$$

Uniqueness of h is not required for identification of the causal functional. Informally, variation in Z shifts the distribution of U within strata of (T, X), the resulting shift in W re-expresses the confounding contribution to O, and if Z did not move the distribution of W given (T, X) the bridge equation would carry no information.

A.3 Linear specialization

Under the linear SEM (Kuroki & Pearl, 2014; Tchetgen Tchetgen et al., 2020)

$$T = \alpha_1 Z + \alpha_2 U + \varepsilon_T, \qquad O = \beta T + \beta_3 U + \beta_4 W + \varepsilon_O, \qquad W = \gamma_1 U + \varepsilon_W, \qquad Z = \delta_1 U + \varepsilon_Z, \tag{10}$$

with mutually independent disturbances independent of U, the ATE equals β; α_1 may be zero (Z a pure proxy), $W \to O$ is allowed, and W must not be caused by (T, Z). Take the linear bridge $h(T, W) = \beta T + \eta W$. The bridge equation equation 8 makes $O - \beta T - \eta W$ uncorrelated with T and with Z, which is two moment equations,

$$\sigma_{TO} = \beta\sigma_{TT} + \eta\sigma_{TW}, \qquad \sigma_{ZO} = \beta\sigma_{ZT} + \eta\sigma_{ZW}. \tag{11}$$

Multiplying the first by σ_{ZW}, the second by σ_{TW} and subtracting eliminates η:

$$\sigma_{TO}\sigma_{ZW} - \sigma_{TW}\sigma_{ZO} = \beta\big(\sigma_{TT}\sigma_{ZW} - \sigma_{TW}\sigma_{TZ}\big) = \beta\,\mathcal{D}, \tag{12}$$

which is equation 1 whenever $\mathcal{D} \neq 0$. Here $\mathcal{D}$ is the determinant of the 2×2 moment system equation 11: the two equations must carry different enough information for β and η to be separated. Because the derivation uses only these two moment equations and never the marginal relation between Z and W, the two proxies enter the formula asymmetrically even when, as in a DNCT, they are structurally interchangeable. The same expression is the population target of proximal two-stage least squares: regress W on (T, Z), then O on $(T, \widehat{W})$; the first-stage coefficient on Z is $\pi_2 = \mathcal{D}/(\sigma_{TT}\sigma_{ZZ} - \sigma_{TZ}^2)$, so $\pi_2 \neq 0$ is the same condition.

A.4 Relevance and denominator stability

For scalar T,

$$\sigma_{ZW\cdot T} = \sigma_{ZW} - \frac{\sigma_{ZT}\sigma_{TW}}{\sigma_{TT}} \qquad \text{and} \qquad \mathcal{D} = \sigma_{TT}\sigma_{ZW} - \sigma_{TW}\sigma_{TZ} = \sigma_{TT}\,\sigma_{ZW\cdot T}, \tag{13}$$

where $\sigma_{ZW\cdot T}$ is the partial covariance of Z and W given T. Since $\sigma_{TT} > 0$, $\mathcal{D} \neq 0$ if and only if $\sigma_{ZW\cdot T} \neq 0$: after linearly removing the treatment, the proxies must still covary, because that residual dependence is what carries the information about the hidden confounding structure. This is the linear form of completeness. Relevance is not binary in finite samples: if $\mathcal{D} = 0.001$, sampling variation can make $\widehat{\mathcal{D}}$ equal 0.0001 or -0.0003, and the ratio estimator explodes. Two properties must therefore be kept apart:

Validity: the structural restrictions of Appendix A.2 hold.

Relevance: the proxies carry enough usable information about U to identify β stably.

A proxy can be valid and still uninformative.

Under the DNCT structure the two are linked by an exact identity. For a disconnected pair every covariance involving Z or W factors through U (Appendix B): $\sigma_{ZW} = \lambda_Z\lambda_W\sigma_U^2$, $\sigma_{TZ} = \lambda_T\lambda_Z\sigma_U^2$, $\sigma_{TW} = \lambda_T\lambda_W\sigma_U^2$, while $\sigma_{TT} = \lambda_T^2\sigma_U^2 + \mathrm{Var}(\varepsilon_T)$. Substituting,

$$\mathcal{D} = \sigma_{TT}\sigma_{ZW} - \sigma_{TW}\sigma_{TZ} = \lambda_Z\lambda_W\sigma_U^2\big(\lambda_T^2\sigma_U^2 + \mathrm{Var}(\varepsilon_T)\big) - \lambda_T^2\lambda_Z\lambda_W\sigma_U^4 = \sigma_{ZW}\,\mathrm{Var}(\varepsilon_T), \tag{14}$$

where $\mathrm{Var}(\varepsilon_T) > 0$ is the part of the treatment not explained by U. So for disconnected negative controls, relevance holds exactly when the two proxies are correlated, that is, when both load on U, in which case each is also correlated with T and with O; and $\mathcal{D}$ is small only when the proxies are weakly correlated or when T is almost a deterministic function of U. The tetrad test checks the structure; the correlation requirement checks the size of the loadings. Candidate proxies are therefore always required to be correlated with one another and with T and O before the tetrad test is applied, and for a triplet with the DNCT structure this requirement is the relevance condition. In the TCGA analysis, candidate pools are correlation-coherent clusters in which every pair of genes, the treatment included, has $|\widehat{r}| \geq 0.35$ (Appendix F). In the simulations every loading on U lies in $[0.8, 1.5]$, so every valid pair has $\mathcal{D} = 4u_Zu_W \geq 2.56$; for invalid triplets, whose members also load on V, $\mathcal{D}$ is not of the form equation 14, and a population gate checks $|\mathcal{D}| \geq 0.05$ for every ordered pair of every triplet before a design is used (Appendix D.2). In finite samples the plug-in denominator $\widehat{\mathcal{D}}$ can still be close to zero by chance, so the estimator discards a pair when $|\widehat{\mathcal{D}}| < 0.01$ (simulations) or when $|\widehat{\mathcal{D}}| < 10^{-4}\big(|\widehat{\sigma}_{TT}\widehat{\sigma}_{ZW}| + |\widehat{\sigma}_{TW}\widehat{\sigma}_{TZ}|\big)$ (TCGA). This is a numerical safeguard against dividing by a noisy near-zero, not a change of estimand.

A.5 PLUG-IN ESTIMATOR

Write $\widehat{\beta}^{\mathrm{PCL}} = \widehat{\mathcal{N}}/\widehat{\mathcal{D}}$ with $\widehat{\mathcal{N}} = \widehat{\sigma}_{TO}\widehat{\sigma}_{ZW} - \widehat{\sigma}_{TW}\widehat{\sigma}_{ZO}$. By the law of large numbers and the continuous mapping theorem, $\widehat{\beta}^{\mathrm{PCL}} \to_P \beta$ when $\mathcal{D} \neq 0$. With $\psi_i = (T_iO_i,\, Z_iW_i,\, T_iW_i,\, Z_iO_i,\, T_i^2,\, T_iZ_i)^\top$, $s = \mathbb{E}[\psi_i]$ and $g(s) = (s_1s_2 - s_3s_4)/(s_5s_2 - s_3s_6)$, the multivariate central limit theorem and the delta method give

$$\sqrt{N}\big(\widehat{\beta}^{\mathrm{PCL}} - \beta\big) \Rightarrow \mathcal{N}\big(0,\ \nabla g(s)^\top \mathrm{Var}(\psi_i)\, \nabla g(s)\big), \tag{15}$$

and every component of $\nabla g(s)$ carries a factor $\mathcal{D}^{-1}$ or $\mathcal{D}^{-2}$, which is the analytic form of the weak-relevance problem above. With covariates X, the same formulas apply to the residuals after linear adjustment for X.

B DNCT STRUCTURE, TETRAD FAITHFULNESS, AND THE JOINT TETRAD TEST

B.1 DISCONNECTED NEGATIVE CONTROLS AND TRIPLETS

Appendix A assumed a valid pair (Z, W) was given. DANCE answers "where do Z and W come from?" by making the structural assumptions *stronger* than general PCL, in exchange for something testable. General proximal theory permits $Z \to T$ and $W \to O$. A *disconnected negative control* (DNC) for (T, O) is a measured variable whose only connection to T, to O and to any other DNC runs through U: in the graph, U is a parent of it and it has no other edge. A DNC therefore satisfies the negative-control-exposure restriction (no route to O other than through U) and the negative-control-outcome restriction (not caused by T or Z) at once, so either of two DNCs can play either role, and both assignments identify β (Appendix B.2). Three consequences are used in the paper: the roles are symmetric, so the six ordered pairs of three DNCs are six estimates of one number; a DNC has no causal effect on O, so its own PCL estimand when placed in the treatment role is zero; and relevance reduces to correlation, $\mathcal{D} = \sigma_{ZW}\mathrm{Var}(\varepsilon_T)$ (Appendix A.4).

A *disconnected negative-control triplet* (DNCT) for (T, O) is a set $\{A, B, C\}$ of three DNCs for (T, O) with the same U. Formally, in a linear acyclic SEM over the measured variables and a single latent U with mutually independent disturbances: (i) U is a parent of each of A, B, C, T, O; (ii) there are no directed edges among A, B, C; (iii) there are no directed edges between any of A, B, C and T or O; (iv) every trek involving at least one of A, B, C passes through U, equivalently each of A, B, C is d-separated from $\{T, O\}$ and from the other two members given U; (v) the edge $T \to O$ is allowed (Figure 4). A DNCT is a structured subclass of the proxies general PCL would accept; the price of testability is that valid-but-connected proxies are excluded by construction. The reason for three rather than two is in Appendix B.4: a pair supplies one testable restriction, a triplet four. For discovery this means that the search is over a stricter class than PCL needs, so some usable proxies are never found (a coverage cost), but everything that is found carries observable restrictions, and two of the DNC properties, symmetry of the roles and no effect on O, can be checked on the accepted set (Section 4).

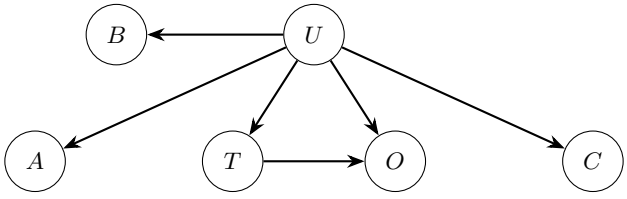


Figure 4: Disconnected negative control triplet $\{A, B, C\}$ for (T, O): every trek between any two of $\{A, B, C, T, O\}$ passes through U; the direct edge $T \to O$ is allowed.

B.2 ONE-FACTOR COVARIANCES

Take everything residualized on X and centred, and write $A = \lambda_A U + \varepsilon_A$, $B = \lambda_B U + \varepsilon_B$, $C = \lambda_C U + \varepsilon_C$, $T = \lambda_T U + \varepsilon_T$, $O = \beta T + \gamma U + \varepsilon_O$, with $\sigma_U^2 = \mathrm{Var}(U)$ and all disturbances mutually independent and independent of U. Substituting the treatment equation into the outcome

gives the reduced form $O = \lambda_O U + \beta\varepsilon_T + \varepsilon_O$ with $\lambda_O = \beta\lambda_T + \gamma$. Independence then gives, for every $j \in \{A, B, C\}$ and every $k \neq j$,

$$\sigma_{jk} = \lambda_j \lambda_k \sigma_U^2, \tag{16}$$

because every cross term such as $\mathrm{Cov}(\varepsilon_A, \varepsilon_B)$ or $\mathrm{Cov}(U, \varepsilon_B)$ vanishes; this product form is the entire reason tetrads work. The one exception is

$$\sigma_{TO} = \lambda_T \lambda_O \sigma_U^2 + \beta \mathrm{Var}(\varepsilon_T) : \tag{17}$$

the treatment disturbance ε_T enters T and then, through β, enters O. Every constraint used for testing avoids σ_{TO}, so the edge being estimated does not enter the test that screens the proxies. Two consequences for estimation. First, both proxy roles are interchangeable: substituting equation 16–equation 17 into equation 1, the numerator is $\beta\mathrm{Var}(\varepsilon_T)\lambda_Z\lambda_W\sigma_U^2$ and the denominator is $\mathrm{Var}(\varepsilon_T)\lambda_Z\lambda_W\sigma_U^2$, so $\beta^{\mathrm{PCL}}(Z, W) = \beta^{\mathrm{PCL}}(W, Z) = \beta$ for any two DNCs with $\mathcal{D} \neq 0$ (in the V-factor population all six ordered assignments of a valid triplet return $\beta = 0.5$ to machine precision, while those of an invalid triplet return six different values, which is why $\bar{\beta}_i$ in Section 2.3 averages over ordered pairs rather than over the three unordered ones). Second, the denominator just computed is $\mathcal{D} = \sigma_{ZW}\mathrm{Var}(\varepsilon_T)$, the relevance identity of Appendix A.4. For discovery this means that everything the screen can see is second-order structure among the candidates, T and O, that the causal edge never enters it, and that requiring candidates to be correlated is the same as requiring the resulting proxy pairs to be relevant.

B.3 Tetrads on four variables: a primer

Four variables $\{A, B, C, D\}$ can be split into two pairs in three ways, giving three tetrads in the notation $T_{ij|k\ell} = \sigma_{ik}\sigma_{j\ell} - \sigma_{i\ell}\sigma_{jk}$ of Section 2.1:

$$T_{AB|CD} = \sigma_{AC}\sigma_{BD} - \sigma_{AD}\sigma_{BC},\ T_{AC|BD} = \sigma_{AB}\sigma_{CD} - \sigma_{AD}\sigma_{CB},\ T_{AD|BC} = \sigma_{AB}\sigma_{DC} - \sigma_{AC}\sigma_{DB}, \tag{18}$$

of which only two are linearly independent, since $T_{AD|BC} = T_{AC|BD} - T_{AB|CD}$ for every covariance matrix. Each is the determinant of the 2×2 block of covariances between one pair and the other. In the one-factor graph of Figure 5(a), where a single latent U is the only common cause of A, B, C, D and there are no edges among them, Wright's path-tracing rules give $\sigma_{jk} = \lambda_j\lambda_k\mathrm{Var}(U)$ for every pair, so every 2×2 block has rank one and all three tetrads vanish, whatever the nonzero loadings. One vanishing tetrad is not diagnostic of this structure. In the fully observed chain $A \to B \to C \to D$ of Figure 5(b), with coefficients a, b, c and unit-variance variables, $\sigma_{AB} = a$, $\sigma_{BC} = b$, $\sigma_{CD} = c$, $\sigma_{AC} = ab$, $\sigma_{BD} = bc$ and $\sigma_{AD} = abc$, so

$$\begin{aligned} &T_{AB|CD} = ab \cdot bc - abc \cdot b = 0, && T_{AC|BD} = ac - abc \cdot b = ac(1 - b^2) \neq 0, \\ &T_{AD|BC} = ac - ab \cdot bc = ac(1 - b^2) \neq 0, \end{aligned} \tag{19}$$

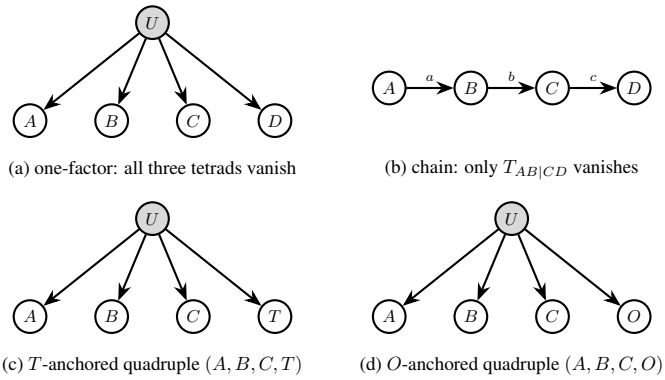
one tetrad vanishes although there is no latent common cause, while the other two do not. Distinguishing the one-factor structure therefore requires the tetrads jointly, which is what the DNCT test does. Under a DNCT the two quadruples (A, B, C, T) and (A, B, C, O) of Figure 5(c, d) are each of the one-factor form (for O, with effective loading $\beta a + c$ after substituting $T = aU + \varepsilon_T$), which is why each contributes three vanishing tetrads; these are the six tetrads used by DANCE and by the joint test, written out in equation 20 below.

B.4 Fifteen tetrads, nine vanishing, six used, four independent

With $T_{ij|k\ell} = \sigma_{ik}\sigma_{j\ell} - \sigma_{i\ell}\sigma_{jk}$, the five variables $\{A, B, C, T, O\}$ have five subsets of four, each with three tetrads: fifteen in all. Under a DNCT nine vanish. The six of the quadruples (A, B, C, T) and (A, B, C, O),

$$\begin{aligned} &T_{AB|CT} = \sigma_{AC}\sigma_{BT} - \sigma_{AT}\sigma_{BC},\ T_{AC|BT} = \sigma_{AB}\sigma_{CT} - \sigma_{AT}\sigma_{CB},\ T_{CB|AT} = \sigma_{CA}\sigma_{BT} - \sigma_{CT}\sigma_{BA}, \\ &T_{AB|CO} = \sigma_{AC}\sigma_{BO} - \sigma_{AO}\sigma_{BC},\ T_{AC|BO} = \sigma_{AB}\sigma_{CO} - \sigma_{AO}\sigma_{CB},\ T_{CB|AO} = \sigma_{CA}\sigma_{BO} - \sigma_{CO}\sigma_{BA}, \end{aligned} \tag{20}$$

the three T-anchored in the first row and the three O-anchored in the second, never contain σ_{TO}; each is the determinant of a block which, under equation 16, is $\sigma_U^2(\lambda_i, \lambda_j)^\top(\lambda_k, \lambda_\ell)$, an outer product of rank one, so it vanishes for every value of the loadings, which is why a vanishing tetrad is



(a) one-factor: all three tetrads vanish

(b) chain: only $T_{AB|CD}$ vanishes

(c) T-anchored quadruple (A, B, C, T)

(d) O-anchored quadruple (A, B, C, O)

Figure 5: **Why tetrads, and which six.** (a) With a single latent common cause and no edges among the observed variables, all three tetrads of equation 18 vanish. (b) A fully observed chain satisfies one tetrad equality but not the other two, so a single vanishing tetrad does not indicate a latent common cause. (c, d) Under a DNCT (Figure 1A) the quadruples (A, B, C, T) and (A, B, C, O) are each of the one-factor form (a), and their three tetrads each are the six tetrads equation 20 tested by DANCE and combined by the joint test.

evidence of one-factor structure and not a numerical accident. Each quadruple containing both T and O, such as (A, B, T, O), contributes one more, $T_{AB|TO} = \sigma_{AT}\sigma_{BO} - \sigma_{AO}\sigma_{BT} = 0$; its other two tetrads contain σ_{TO} and equal $\sigma_{AB}\sigma_{TO} - \sigma_{AO}\sigma_{TB} = \lambda_A\lambda_B\beta\sigma_U^2\mathrm{Var}(\varepsilon_T) \neq 0$, the common-factor pieces cancelling and the causal term remaining. This is also why a pair is untestable: of the three tetrads of (Z, W, T, O), only $T_{ZW|TO}$ vanishes. DANCE uses the six that contain all three proxies. The three pair-type constraints are not lost: from $T_{AB|CT} = 0$ and $T_{AB|CO} = 0$, $\sigma_{AT}/\sigma_{BT} = \sigma_{AC}/\sigma_{BC} = \sigma_{AO}/\sigma_{BO}$ whenever $\sigma_{BC} \neq 0$, hence $\sigma_{AT}\sigma_{BO} = \sigma_{AO}\sigma_{BT}$, which is $T_{AB|TO} = 0$; so on any triplet with nonzero proxy–proxy covariances, the relevance requirement, the six imply the nine. Within each anchored family the three tetrads satisfy one exact linear relation,

$$T_{CB|AT} = T_{AB|CT} - T_{AC|BT}, \qquad T_{CB|AO} = T_{AB|CO} - T_{AC|BO} : \tag{21}$$

expanding, $(\sigma_{AC}\sigma_{BT} - \sigma_{AT}\sigma_{BC}) - (\sigma_{AB}\sigma_{CT} - \sigma_{AT}\sigma_{CB}) = \sigma_{AC}\sigma_{BT} - \sigma_{AB}\sigma_{CT}$, the two σ_{AT} terms cancelling because $\sigma_{BC} = \sigma_{CB}$; identically with O in place of T. The relation holds at every Σ, not only under the null, so the six constraints have algebraic rank four and the testable null is the four-vector $t(\Sigma_i)$ of Section 2.1; the count matches the parameters, five loadings and σ_{TO} inside ten off-diagonal covariances. Algebraic independence is not statistical independence: $\widehat{T}_{AB|CT}$ and $\widehat{T}_{AC|BT}$ share the term $-\widehat{\sigma}_{AT}\widehat{\sigma}_{BC}$ outright, so sampling error in one covariance moves both, and in a typical simulated triplet the pairwise correlations of the four tetrads are 0.30–0.76 (Appendix B.6). For discovery this fixes three things: a single pair cannot be screened at all; every candidate triplet carries exactly four degrees of freedom of evidence, which sets the reference distribution of the test; and those four pieces of evidence are correlated, so they must be combined through their joint covariance rather than as four separate tests.

B.5 Structural validity versus tetrad compatibility

Appendices B.2–B.4 prove DNCT $\Rightarrow \boldsymbol{t}(\Sigma) = \mathbf{0}$. The converse is *tetrad faithfulness*: a tetrad vanishes only because of the structural rank deficiency, never through a coincidence of parameter values. The coincidence exists. If a second latent factor V, independent of U, loads on the three candidates in exactly the same proportion as U, $v_j = c\,u_j$, then each candidate is $u_j(U + cV) +$ noise: from their point of view there is one latent direction, every covariance involving a candidate keeps the product form equation 16 with effective loadings $u_j(\sigma_U^2 + c^2\sigma_V^2)^{1/2}$, and all nine tetrads are exactly zero at every N. If V also drives O, the triplet is invalid and its PCL estimand is biased by an amount that does not shrink with N (closed form in Appendix D.2). No sample size and no multiple-testing machinery separates such a triplet from a valid one: this is an identification problem, not a power problem. Faithfulness assumes that exact proportionality does not occur, so that $H_i = 1 \Rightarrow \boldsymbol{t}(\Sigma_i) \neq \mathbf{0}$ and hence $\boldsymbol{t}(\Sigma_i) = \mathbf{0} \iff H_i = 0$. It is the same kind of genericity assumption as faithfulness in constraint-based causal discovery, where a conditional independence is taken to reflect the graph

rather than a cancellation of path coefficients; it fails only on a measure-zero set of parameter values and is untestable from data. A near-miss, proportions almost but not exactly equal, is a power problem rather than a faithfulness failure: the tetrads are nonzero but small, and detectability is a matter of N, which is what the near-null regime of the simulations represents.

Two levels, and what π_0 measures. The structural truth of a candidate is H_i; the observable null is $\boldsymbol{t}(\Sigma_i) = \mathbf{0}$. Everything the paper estimates and controls is indexed to the observable null. In particular the uniform component of the fitted mixture is the set of tetrad-compatible candidates, not the set of valid ones: if 20% of candidates are true DNCTs, 10% are invalid but exactly tetrad-compatible and 70% are invalid and tetrad-violating, the first two groups both produce uniform p-values and the observed mixture has a uniform component of 0.30 while the valid proportion is 0.20, at any N. Reading $\widehat{\pi}_0$ as the prevalence of validity and the FVR as structural contamination requires faithfulness on top of the statistics. Two failure modes must not be confused: a tetrad-compatible invalid triplet ($H_i = 1$, $\boldsymbol{t} = \mathbf{0}$) is a faithfulness failure, whereas a harmless invalid triplet ($H_i = 1$, $\boldsymbol{t} \neq \mathbf{0}$, $b_i \approx 0$) is not; its tetrads are visibly nonzero and parameter cancellation merely makes its estimand unbiased. Invalidity does not imply harm, which is why severity is recorded separately from validity.

Several hidden variables. By trek separation (Sullivant et al., 2010) a tetrad vanishes generically if and only if all treks behind its four covariances pass through a common choke point on the same side of every trek. The three T-anchored tetrads force one bottleneck for every trek from A, B, C into T, the O-anchored three do the same for O, and because $T \to O$ the two bottlenecks coincide: the six constraints hold generically only if A, B, C, T, O share a single *effective* latent factor, one confounder or several on one non-colliding path, which induce the same one-dimensional covariance structure and are indistinguishable from a single U. Two confounders with linearly independent covariance patterns generically violate at least one tetrad, and the pipeline rejects such triplets at the cost of discarding usable proxies.

For discovery this fixes what the pipeline can and cannot promise: its guarantees concern tetrad compatibility; the structural reading is an assumption; the blind spot is invisible at every sample size, so the simulation design keeps every panel off it deliberately (evenly spaced loading ratios, Appendix D.2) and the near-null regime approaches it as a power problem; and only an effective one-factor confounder is discoverable, which is why the real-data enumeration is restricted to correlation-coherent clusters.

B.6 From the population constraint to the test

We never observe Σ. From N units we form $\widehat{\Sigma}$ and $\widehat{\boldsymbol{t}} = \boldsymbol{t}(\widehat{\Sigma})$, and even under an exact DNCT $\widehat{\boldsymbol{t}}$ fluctuates around zero; the question is whether the observed $\widehat{\boldsymbol{t}}$ is plausibly sampling noise around zero. A single covariance is a sample mean of products, so $\sqrt{N}(\widehat{\sigma}_{AB} - \sigma_{AB})$ is asymptotically normal; the tetrads use several covariances from the same units ($\widehat{\sigma}_{AC}$ and $\widehat{\sigma}_{AT}$ share every A_r), so their errors are correlated and the argument must be multivariate.

Central limit theorem for the sample covariances. Collect the ten off-diagonal covariances of (A, B, C, T, O) in $\widehat{\mathbf{s}}$. Under finite fourth moments, $\sqrt{N}(\widehat{\mathbf{s}} - \mathbf{s}) \Rightarrow \mathcal{N}(\mathbf{0}, \Sigma_{ss})$. Under multivariate normality Isserlis' theorem gives the entries in closed form,

$$(\Sigma_{ss})_{(ij),(k\ell)} = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}; \tag{22}$$

the limit theorem is general, the closed form is Gaussian-specific (Appendix B.9).

Delta method. Each tetrad depends on four covariances, with $\partial T_{ij|k\ell}/\partial\sigma_{ik} = \sigma_{j\ell}$, $\partial/\partial\sigma_{j\ell} = \sigma_{ik}$, $\partial/\partial\sigma_{i\ell} = -\sigma_{jk}$, $\partial/\partial\sigma_{jk} = -\sigma_{i\ell}$. With $\mathbf{s} = (\sigma_{AB}, \sigma_{AC}, \sigma_{AT}, \sigma_{AO}, \sigma_{BC}, \sigma_{BT}, \sigma_{BO}, \sigma_{CT}, \sigma_{CO}, \sigma_{TO})^\top$, the 4×10 Jacobian $D = \partial\boldsymbol{t}/\partial\mathbf{s}$ has rows

$$\begin{aligned}
\nabla T_{AB|CT}^\top &= (\,0,\ \sigma_{BT},\ -\sigma_{BC},\ 0,\ -\sigma_{AT},\ \sigma_{AC},\ 0,\ 0,\ 0,\ 0\,),\\
\nabla T_{AC|BT}^\top &= (\,\sigma_{CT},\ 0,\ -\sigma_{BC},\ 0,\ -\sigma_{AT},\ 0,\ 0,\ \sigma_{AB},\ 0,\ 0\,),\\
\nabla T_{AB|CO}^\top &= (\,0,\ \sigma_{BO},\ 0,\ -\sigma_{BC},\ -\sigma_{AO},\ 0,\ \sigma_{AC},\ 0,\ 0,\ 0\,),\\
\nabla T_{AC|BO}^\top &= (\,\sigma_{CO},\ 0,\ 0,\ -\sigma_{BC},\ -\sigma_{AO},\ 0,\ 0,\ 0,\ \sigma_{AB},\ 0\,),
\end{aligned} \tag{23}$$

and the σ_{TO} column is identically zero: the causal edge does not enter the sampling behaviour of the test. Then $\sqrt{N}\{\boldsymbol{t}(\widehat{\Sigma}) - \boldsymbol{t}(\Sigma)\} \Rightarrow \mathcal{N}(\mathbf{0}, \Sigma_t)$ with $\Sigma_t = D\Sigma_{ss}D^\top$, evaluated at $\widehat{\Sigma}$. Σ_t records how noisy each tetrad is and how the four move together (Figure 6A–B).

Wald statistic and null distribution. Under $\mathcal{H}_0^{\text{tet}}$, $\sqrt{N}\widehat{\boldsymbol{t}} \Rightarrow \mathcal{N}(\mathbf{0}, \Sigma_t)$. The squared Mahalanobis distance $N\widehat{\boldsymbol{t}}^\top\Sigma_t^{-1}\widehat{\boldsymbol{t}}$ whitens the vector ($Z = \Sigma_t^{-1/2}\sqrt{N}\widehat{\boldsymbol{t}} \sim \mathcal{N}(\mathbf{0}, I_4)$) and sums four squared standard normals, hence χ^2_4; if Σ_t has rank $q < 4$ the Moore–Penrose inverse gives χ^2_q. This is equation 2; Figure 6C shows the simulated null law against χ^2_4 (mean 4.02, realized size 0.043 at nominal 0.05), and Figure 8B the resulting uniform p-values (null QQ plot).

Behaviour under alternatives. Under a fixed alternative $\boldsymbol{t}(\Sigma) \neq \mathbf{0}$ the quantity $N\boldsymbol{t}^\top\Sigma_t^+\boldsymbol{t}$ grows with N, so $T_i \to \infty$ at rate N and the test is consistent; the noncentral $\chi^2_q(\nu)$ with $\nu = N\boldsymbol{t}^\top\Sigma_t^+\boldsymbol{t}$ is the finite-N approximation, exact as a weak limit only along $N^{-1/2}$-local alternatives (Figure 6E). In that approximation the whole effect of a violation on the test is the one number ν, linear in N and in the squared size of the violation. For the five simulation regimes at $N = 2000$ the representative noncentralities are 5.7×10^{-18}, 2.05, 45.6, 167 and 236 (near null to large): the near-null regime is numerically the null, and the very-small regime is where impurity (Proposition 2) bites.

Effect of the tetrad covariance on calibration. The four tetrads of a valid triplet are positively correlated: 0.30–0.76 in the population of Table 1 (Figure 6B). A statistic that standardizes each tetrad by its own variance and sums the squares, $\sum_j \widehat{t}_j^2/(\widehat{\Sigma}_t)_{jj}$, has mean 4 like a χ^2_4 but variance 13–15 instead of 8, so referred to χ^2_4 its size at nominal 0.05 is about 0.08 (Table 1, Figure 6D). When two proxies are nearly duplicates the smallest eigenvalue of Σ_t falls from 6.6×10^{-3} to 2.0×10^{-5} and the eigenvalue spread widens from a factor 31 to 1330; with the ridge and pseudo-inverse below the realized size in that design is 0.045 (Figure 6F). For discovery this is the property the acceptance layer stands on: the two-group model needs $f_0 \equiv 1$, and a per-tetrad or naively combined statistic would feed it a miscalibrated null.

Table 1: Joint statistic against the sum of squared standardized tetrads. One valid-triplet population with loadings 0.9, 1.15, 1.42; $N = 2000$; 3000 replicates per seed; size is the fraction of replicates with $p \leq 0.05$. Reference values for χ^2_4: mean 4, variance 8.

seed	joint statistic, size at 5%	naive sum, size at 5%	naive sum, mean	naive sum, variance
1 (Figure 6)	0.043	0.081	4.03	14.2
7	0.043	0.081	3.96	13.3
11	0.050	0.086	4.02	15.3

Implementation. $\widehat{\Sigma}_t$ is built from the sample covariance with equation 22, a ridge of 10^{-7} is added to its diagonal, q is the numerical rank at absolute tolerance 10^{-10}, the quadratic form uses the pseudo-inverse, and the p-value is clipped at zero. Because the ridge exceeds the rank tolerance, $q = 4$ whenever the four tetrads are computable: in all 5,710,820 TCGA tests and in every simulation cell. Simulations and TCGA use the same function. (Writing the asymptotic covariance with an explicit N in the quadratic form, as here, or $\mathrm{Var}(\widehat{\boldsymbol{t}}) \approx \Sigma_t/N$ without it, as the code does, gives the same number.)

B.7 The two Wishart comparators

For one tetrad $D = T_{ab|cd}$, DANCE uses Wishart's normal-theory variance (Wishart, 1928; Kummerfeld et al., 2024)

$$\widehat{\mathrm{Var}}(\widehat{D}) = \frac{D_{ab}D_{cd}\,(N+1)/(N-1) - D_{abcd}}{N-2}, \qquad D_{ab} = \widehat{\sigma}^2_{ab} - \widehat{\sigma}_{aa}\widehat{\sigma}_{bb}, \quad D_{cd} = \widehat{\sigma}^2_{cd} - \widehat{\sigma}_{cc}\widehat{\sigma}_{dd}, \tag{24}$$

with D_{abcd} the determinant of the 4×4 sample covariance of (a, b, c, d), and a two-sided normal p-value for $\widehat{D}/\widehat{\mathrm{Var}}(\widehat{D})^{1/2}$; each triplet contributes six component p-values. *DANCE all-pass* accepts a triplet if and only if the smallest of its six p-values exceeds $\alpha = 1/N$, the published default; no

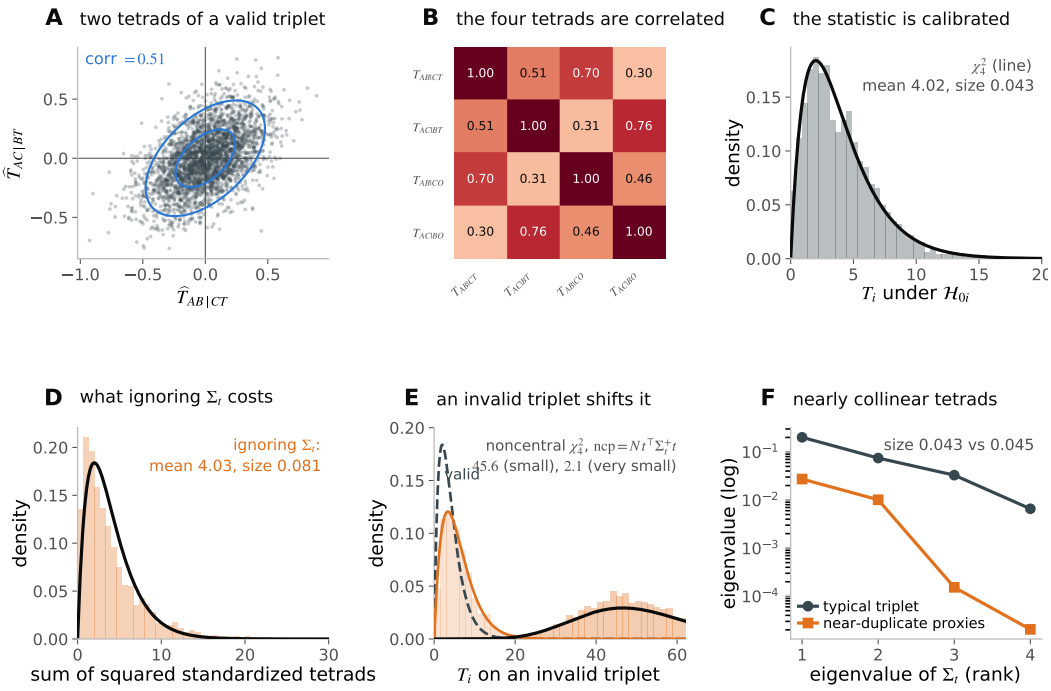


Figure 6: From four tetrads to one p-value; $N = 2000$, 3000 replicates, the paper's test code. **A:** two of the four sample tetrads of a valid triplet with the one- and two-standard-deviation ellipses implied by Σ_t. **B:** correlation matrix of the four tetrads. **C:** the joint statistic against its χ^2_4 reference. **D:** the sum of squared standardized tetrads, which ignores the off-diagonal entries of Σ_t: same mean, heavier tails, size 0.081. **E:** the statistic on invalid triplets of the very-small and small regimes against the noncentral law, $\nu = 2.05$ and 45.6 (the panel label rounds the first to 2.1). **F:** eigenvalues of Σ_t for a typical triplet and for a design with near-duplicate proxies.

mixture is fitted. *Wishart-component* applies the local-validity procedure of Section 2.2 to the same six tests: the mixture is fitted to all $6m$ component p-values and a triplet is accepted if and only if all six component local validities reach τ. It differs from the joint test in two ways only, the covariance among a triplet's tetrads is ignored and two of the six tetrads are redundant, so the comparison isolates the statistic and the acceptance geometry (all six versus one joint) rather than the threshold. The dashed Wishart-component curves of Figure 8C report $\widehat{\pi}_0$ from the mixture fitted to the $6m$ per-tetrad p-values; DANCE all-pass fits no mixture and has no $\widehat{\pi}_0$. In pure designs the tetrad-level null proportion equals the triplet-level π_0; in mixed designs it is $(6m_0 + 4m_1 + m_2)/(6\sum_K m_K)$, because a triplet with one violation gene leaves the three T-anchored tetrads and one O-anchored tetrad exactly zero and a triplet with two leaves one (Appendix D.2).

B.8 Within- versus across-triplet dependence

$\widehat{\Sigma}_t$ accounts for the dependence among the four tetrads of one candidate, which is what makes one joint p-value coherent. It does nothing about dependence across candidates: triplets that share genes share entries of $\widehat{\Sigma}$, and in a full enumeration every gene appears in $\binom{p-1}{2}$ triplets. That dependence remains in $\{P_i\}$ and is studied in Section 3.2 and Appendix D.3; for discovery it means that m nominal candidates carry far fewer than m independent pieces of information about the mixture.

B.9 Non-normal data and nonlinearity

For non-normal data Σ_{ss} acquires fourth-moment terms and equation 22 can misscale the statistic. A Satorra–Bentler-type sandwich estimate from sample fourth moments (Satorra & Bentler, 1994) or a permutation reference distribution can replace the normal-theory covariance; neither is implemented

in the released code, whose simulations are Gaussian by construction and whose expression data are log-scale values used with the normal-theory covariance as is. Nonlinearity in the latent relation is a different failure, under which even the population tetrads of a valid triplet are nonzero; the generalized tetrad test of Ying et al. (2025) replaces the covariance constraint, and the acceptance layer is unchanged (Appendix E.4). For discovery both are calibration risks: a miscalibrated null perturbs $\widehat{\pi}_0$ and $\widehat{L}$ jointly, which is why the simulations evaluate the fitted rule directly rather than reasoning from the sign of one error.

C Acceptance-side validation: derivations, proofs and worked examples

This appendix supports Section 2.2. Throughout, $H_i = 0$ marks a valid triplet, $H_i = 1$ an invalid one, $f_0 \equiv 1$ is the density of a valid triplet's p-value and f_1 that of an invalid triplet's, $\bar{F}_1(\gamma) = \int_\gamma^1 f_1$, and $q = 4$.

C.1 Calibration and the shape of the invalid density

A valid triplet is tetrad-compatible, so its Wald statistic T is asymptotically χ^2_4 (Appendix B.6) and $P = 1 - F_4(T)$ is Uniform$(0, 1)$ by the probability integral transform: $\mathbb{P}(P \le p) = \mathbb{P}(T \ge F_4^{-1}(1-p)) = p$. Figure 8B and Table 1 verify this at finite N; the only departure is slight conservativeness at $N = 500$ (size 0.041). For an invalid triplet, T is asymptotically noncentral $\chi^2_4(\nu_i)$ with $\nu_i = N\, t_i^\top \Sigma^+_{t,i} t_i$ (Appendix B.6). The noncentral law is a Poisson$(\nu/2)$ mixture of central χ^2 laws with $4 + 2K$ degrees of freedom, each stochastically larger than χ^2_4, so T is stochastically larger under $H_i = 1$ and P is stochastically smaller: $F_1(p) = \mathbb{P}(P \le p \mid H_i = 1) \ge p$, strictly when $\nu > 0$. The invalid p-value density is the likelihood ratio at the observed statistic,

$$f_1(p) = \frac{g_\nu(t_p)}{g_4(t_p)}, \qquad t_p = F_4^{-1}(1-p), \tag{25}$$

with g_ν the noncentral and g_4 the central density; it is decreasing in p and tends to $e^{-\nu/2}$ as $p \to 1$ (proof in Appendix C.8). Table 2 makes the shape concrete: at the very-small-regime noncentrality the invalid density is a gentle slope over the whole interval, not a spike at zero, and only from the small regime onward does it vanish near one. Figure 2B shows the fitted densities, and Figure 8A shows them regime by regime.

Table 2: Invalid p-values under the noncentral $\chi^2_4(\nu)$ law, for the representative noncentralities of Appendix B.6: the probability that $P \le p$, and the density at the right end, $f_1(1) = e^{-\nu/2}$.

regime (cell)	ν	$\mathbb{P}(P \le 0.05)$	$\mathbb{P}(P \le 0.2)$	$\mathbb{P}(P \le 0.5)$	$\mathbb{P}(P \le 0.8)$	$f_1(1)$
very small ($N = 2000$)	2.05	0.175	0.42	0.715	0.907	0.36
small ($N = 500$)	11.4	0.78	0.93	0.985	0.998	0.003
small ($N = 2000$)	45.6	1.000	1.000	1.000	1.000	10^{-10}

For discovery: the p-values carry class information only through the gap between f_1 and the flat f_0; this gap, not the number of candidates, is the resource every later step spends. In the near-null regime ($\nu \approx 10^{-17}$) $f_1 = f_0$ to numerical precision, so no rule applied to those p-values can separate the classes (Figure 2B–D). That is a power problem and is distinct from the rank-one blind spot of Appendix B.5, where the tetrads vanish exactly while the bias does not.

C.2 What a fixed raw-p cutoff controls

Bayes' rule gives $\mathrm{FVR}(\gamma) = \mathbb{P}(H_i = 1 \mid P > \gamma) = \mathbb{P}(P > \gamma \mid H_i = 1)\mathbb{P}(H_i = 1)/\mathbb{P}(P > \gamma)$. The numerator is $(1 - \pi_0)\bar{F}_1(\gamma)$, and $\mathbb{P}(P > \gamma) = \pi_0(1 - \gamma) + (1 - \pi_0)\bar{F}_1(\gamma)$ is the sum of the accepted valid and the accepted invalid mass, which gives equation 4; dividing both integrands in the last form of equation 4 by f_P gives $\mathrm{FVR}(\mathcal{R}) = \mathbb{E}[1 - L(P) \mid P \in \mathcal{R}]$ for $\mathcal{R} = (\gamma, 1]$ (Lemma 1). Three quantities appear, γ, π_0 and $\bar{F}_1(\gamma)$; the analyst fixes γ and the data-generating process fixes the other two. Table 3 evaluates equation 4 at $\gamma = 0.05$ in three settings.

Table 3: Contamination of the rule "accept $P > 0.05$" from equation 4. Valid pile $= \pi_0 \cdot 0.95$, invalid pile $= (1-\pi_0)\bar{F}_1(0.05)$.

π_0	$\bar{F}_1(0.05)$	valid pile	invalid pile	FVR(0.05)	reading
0.9	0.20	0.855	0.020	0.023	valid common, violations detectable
0.1	0.20	0.095	0.180	0.655	same test, same power, valid rare
0.2	0.80	0.190	0.640	0.771	weak violations pass the cutoff

Monotonicity. With $a = 1-\gamma$ and $b = \bar{F}_1(\gamma)$,

$$\frac{\partial \,\mathrm{FVR}}{\partial \pi_0} = \frac{-ab}{\{\pi_0 a + (1-\pi_0)b\}^2} \le 0, \qquad \frac{\partial \,\mathrm{FVR}}{\partial b} = \frac{\pi_0(1-\pi_0)a}{\{\pi_0 a + (1-\pi_0)b\}^2} \ge 0 :$$

growing π_0 grows the valid pile, growing b grows the invalid pile.

For discovery: a data-driven search sits at the bad end of both derivatives. Valid triplets are rare (in TCGA $\widehat{\pi}_0 = 0.157$, Section 4) and weak violations pass any cutoff. This is why DANCE all-pass, which keeps a triplet when all six per-tetrad p-values exceed $1/N$, is heavily contaminated in the weak regimes (0.73 at $N = 1000$ and still 0.33 at $N = 10^4$ in the small regime; Figure 8D, Appendix D.4). In TCGA the same rule retains 3,198,500 of the 5,710,820 candidates (56%), where the posterior rule at $\tau = 0.9$ retains 45,270 (0.8%; Appendix F). A fixed cutoff has no handle that responds to π_0 or to $\bar{F}_1$; Section 2.2 gives it one.

C.3 Local validity: derivation and reading

For $dp \to 0$,

$$\mathbb{P}(H_i = 0 \mid P \in [p, p+dp]) = \frac{\mathbb{P}(P \in [p, p+dp] \mid H_i = 0)\,\pi_0}{\mathbb{P}(P \in [p, p+dp])} = \frac{f_0(p)\,dp\,\pi_0}{f_P(p)\,dp} = \frac{\pi_0}{f_P(p)},$$

since $f_0 \equiv 1$; this is equation 5. The dp cancels: L is a density ratio, not a tail-area ratio, which is why it can differ sharply between neighbouring p-values while $\mathrm{FVR}(\gamma)$ cannot.

Why large L means valid. $L(p) = \pi_0/\{\pi_0 + (1-\pi_0)f_1(p)\}$ equals 1 exactly where $f_1(p) = 0$ and π_0 where $f_1(p) = 1$. It does not measure how large p is; it measures how empty of invalid triplets the neighbourhood of p is. If $f_1 \equiv f_0$ (no information) then $L \equiv \pi_0$ at every p: the p-value adds nothing to the prior, and no threshold above π_0 accepts anything, which is the correct behaviour (Figure 2D, near-null regime).

A two-world map. Take $f_1 = \mathrm{Beta}(1,3)$, density $3(1-p)^2$, in world A with $\pi_0 = 0.9$ and world B with $\pi_0 = 0.1$. Table 4 lists $L(p)$ in both. A p-value of 0.8 is the same number in both worlds but means "99% valid" in A and "48% valid" in B. Table 5 compares rules. The raw rule $P > 0.05$ keeps 95% of valid triplets in both worlds and has contamination 0.091 in A and 0.890 in B, from equation 4 with $\bar{F}_1(0.05) = 0.95^3 = 0.857$. The rule $L \ge 0.9$ has implied cutoffs $p \ge 0.42$ in A and $p \ge 0.94$ in B, from solving $(1-c)^2 = 0.1\pi_0/\{2.7(1-\pi_0)\}$, contamination $1 - \pi_0(1-c)/\{\pi_0(1-c) + (1-\pi_0)(1-c)^3\} = 0.036$ in both, and coverage $1-c$, that is 0.58 in A and 0.06 in B. Contamination is what the posterior rule holds fixed; coverage is what adapts.

Table 4: Local validity $L(p) = \pi_0/\{\pi_0 + (1-\pi_0)\,3(1-p)^2\}$ in two worlds that share $f_1 = \mathrm{Beta}(1,3)$.

p	0.01	0.05	0.2	0.5	0.8	0.95
world A ($\pi_0 = 0.9$)	0.75	0.77	0.82	0.92	0.99	0.999
world B ($\pi_0 = 0.1$)	0.036	0.039	0.055	0.13	0.48	0.94

Relation to existing quantities. L is the local false discovery rate of Efron et al. (2001); see Efron (2010). Thresholding L rather than a tail-area quantity (the acceptance-side analogue of a q-value, Storey, 2003) gives each accepted triplet an individual guarantee, which matters because each accepted triplet is used individually in PCL.

Table 5: Two rules in the two worlds of Table 4: contamination (FVR) and valid-triplet coverage.

	rule $P > 0.05$		rule $L \geq 0.9$		
	FVR	coverage	implied cutoff	FVR	coverage
world A ($\pi_0 = 0.9$)	0.091	0.95	$p \geq 0.42$	0.036	0.58
world B ($\pi_0 = 0.1$)	0.890	0.95	$p \geq 0.94$	0.036	0.06

C.4 Proof of Proposition 1, and population versus realized contamination

Lemma 1 (FVR as an average of $1 - L$)**.** *For any acceptance region $\mathcal{R}$ determined by P alone with $\mathbb{P}(P \in \mathcal{R}) > 0$, $\mathbb{P}(H = 0 \mid P \in \mathcal{R}) = \mathbb{E}[L(P) \mid P \in \mathcal{R}]$, that is,* $\mathrm{FVR}(\mathcal{R}) = \mathbb{E}[1 - L(P) \mid P \in \mathcal{R}]$.

Proof. $\mathbb{E}[\mathbf{1}\{H = 0\}\mathbf{1}\{P \in \mathcal{R}\}] = \mathbb{E}\big[\mathbb{E}[\mathbf{1}\{H = 0\} \mid P]\,\mathbf{1}\{P \in \mathcal{R}\}\big] = \mathbb{E}[L(P)\mathbf{1}\{P \in \mathcal{R}\}]$; divide by $\mathbb{P}(P \in \mathcal{R})$. For $\mathcal{R} = (\gamma, 1]$ this is the last form of equation 4 with both integrands divided by f_P. □

Proof of Proposition 1. Take $\mathcal{R} = \{p : L(p) \geq \tau\}$. Then $\mathrm{FVR}(\mathcal{A}^*_\tau) = 1 - \mathbb{P}(H = 0 \mid P \in \mathcal{R}) = 1 - \mathbb{E}[L(P) \mid L(P) \geq \tau] \leq 1 - \tau$, since $L \geq \tau$ on $\mathcal{R}$. □

Equality holds if and only if $L(P) = \tau$ almost surely on $\mathcal{R}$; with a continuous decreasing f_1 the accepted L values spread over $[\tau, 1]$, so the realized contamination is typically below $1-\tau$ (Figure 2D, moderate and large regimes).

Population versus realized. The proposition bounds a probability for a randomly drawn candidate. In replicate r a rule accepts R_r triplets, I_r of them invalid, and the simulations report I_r/R_r. If (H_i, P_i) are i.i.d. across candidates, then conditional on the accepted p-values the labels of the accepted triplets are independent Bernoulli$(1 - L(P_i))$, so $\mathbb{E}[I_r/R_r \mid R_r > 0,\ \text{accepted } p\text{-values}]$ is the mean of $1 - L(P_i)$ over the accepted, which is $\leq 1 - \tau$. Triplets that share genes are not independent: an extreme case is m identical copies of one triplet, for which $I/R \in \{0, 1\}$ even though its expectation is still $\leq 1 - \tau$. Under dependence the marginal statement remains true, but I_r/R_r is more variable around it (Figure 7B; Appendix C.9).

Abstention. If no candidate reaches τ then $\widehat{\mathcal{A}}_\tau$ is empty, $R_r = 0$, and I_r/R_r is undefined. The simulations count this as an abstention and report its probability separately (Figure 8G; Appendix D.4), never as zero contamination; averaging it in as zero would flatter every rule in exactly the regimes where it cannot decide.

C.5 Transport of a threshold

If f_1 is nonincreasing then L is nondecreasing and $\{L(P) \geq \tau\} = \{P \geq c_\tau\}$ with $L(c_\tau) = \tau$; solving $\pi_0/\{\pi_0 + (1 - \pi_0)f_1(c_\tau)\} = \tau$ gives the transport equation of Section 2.2, $f_1(c_\tau) = \pi_0(1 - \tau)/\{\tau(1 - \pi_0)\}$. For $\tau = 0.9$ and $\pi_0 = 0.191$ the invalid density must have fallen to 0.026 at the cutoff; for the Beta$(1, 3)$ alternative with $\pi_0 = 0.9$, to 1.0.

Two knobs. Both rules accept a p-upper set. The cutoff rule holds the set fixed and lets $\mathrm{FVR}(\gamma)$ move with (π_0, f_1) as Appendix C.2 shows. The posterior rule holds the target fixed; its implied cutoff $\hat{c}_\tau$ is re-solved from each fitted mixture and moves toward one as π_0 falls or as f_1 flattens, until it reaches one and the rule abstains. In the two worlds of Table 5, the same $\tau = 0.9$ produces cutoffs 0.42 and 0.94 and the same 3.6% contamination, at the price of coverage 0.58 versus 0.06.

In the simulations. Figure 2E holds f_0 and f_1 fixed and varies π_0 and N in paired lists of $m = 2000$ p-values (design in Appendix D): the raw-p cutoff $c^*_{0.10}$ that attains FVR 0.10 moves by more than two orders of magnitude with N, shifts again with π_0, and does not exist in the weakest cells (up-triangles); a cutoff $c = 0.075$ tuned on the identifiable cells keeps coverage near 0.93 but yields contamination from about 0 to 0.7 on held-out cells. The threshold $\tau = 0.9$ keeps the meaning "90% posterior validity" in every cell. The advantage is calibration transfer, not ranking.

For discovery: a threshold that transports is the only kind one can set before seeing the data, which is the situation in every new dataset; a cutoff that must be re-tuned per dataset presupposes the labels one is trying to discover.

C.6 Estimation with qvalue

Storey's estimator. $\widehat{\pi}_0(\lambda) = \#\{P_i > \lambda\}/\{m(1-\lambda)\}$. Above λ the valid triplets contribute a fraction $\pi_0(1-\lambda)$ of all p-values and the invalid ones $(1-\pi_0)\bar{F}_1(\lambda)$, so

$$\mathbb{E}\,\widehat{\pi}_0(\lambda) = \pi_0 + (1-\pi_0)\frac{\bar{F}_1(\lambda)}{1-\lambda} \;\geq\; \pi_0, \tag{26}$$

with equality only if no invalid p-value exceeds λ. Increasing λ shrinks the bias ($\bar{F}_1(\lambda)/(1-\lambda) \to \theta$ of Proposition 2) but grows the variance, because fewer p-values lie above λ. Following Storey & Tibshirani (2003, Remark B), `qvalue` computes $\widehat{\pi}_0(\lambda)$ on a grid of λ, fits a natural cubic smoothing spline with three degrees of freedom through the values and reads off the fitted value at the largest grid point; the package defaults, used here, are $\lambda = 0.05, 0.10, \ldots, 0.95$, read at $\lambda = 0.95$ (Storey et al., 2023), whereas the remark in the paper uses $\lambda = 0, 0.01, \ldots, 0.95$ read at $\lambda = 1$.

Density and score. $\widehat{f}_P$ is a Gaussian kernel density estimate on the probit scale (bandwidth adjustment 1.5), transformed back to $[0,1]$. $\widehat{L}(p) = \min\{1, \widehat{\pi}_0/\widehat{f}_P(p)\}$ is then made nondecreasing in p by a running maximum from $p = 0$ upward; $\widehat{L}$ is exactly `qvalue`'s local FDR output.

Failed fits. When the spline is undefined or $\widehat{\pi}_0 \leq 0$ the fit fails; the replicate is recorded as such (Figure 7, caption) and never counted as an accepted set.

For discovery: $\widehat{L}$ is a function of two estimates, and the one that sets the level of every score is $\widehat{\pi}_0$; Appendix C.7 says how an error in it propagates and Appendix C.8 which direction it takes.

C.7 The plug-in bound

Lemma 2 (Plug-in bound)**.** *Let $\widehat{L}$ be a deterministic score function, or a score function fitted on data independent of (P, H), with the statement then holding conditionally on $\widehat{L}$. If $\sup_p |\widehat{L}(p) - L(p)| \leq \varepsilon$ and $\mathbb{P}(\widehat{L}(P) \geq \tau) > 0$, then $\mathrm{FVR}(\widehat{\mathcal{A}}_\tau) \leq 1 - \tau + \varepsilon$.*

On $\{\widehat{L}(P) \geq \tau\}$ we have $L(P) \geq \tau - \varepsilon$, and Lemma 1 with $\mathcal{R} = \{p : \widehat{L}(p) \geq \tau\}$, a region determined by p once $\widehat{L}$ is fixed, gives $1 - \mathrm{FVR} = \mathbb{E}[L(P) \mid P \in \mathcal{R}] \geq \tau - \varepsilon$. In the pipeline $\widehat{L}$ is fitted to the same m candidate p-values that it then thresholds. Under independent candidates the conditioning is innocuous, because H_i is conditionally independent of the other candidates given P_i; under the overlapping-triplet dependence of Section 3.2 it is not automatic, and the same-data setting is therefore assessed empirically through the realized contamination of the simulations rather than covered by the lemma. Example: $\tau = 0.9$ and $\varepsilon = 0.03$ give a guarantee of 0.13, not 0.10.

What ε contains. Sampling error of $\widehat{\pi}_0$ and $\widehat{f}_P$, which shrinks with m_{eff} (Appendix C.9), and the systematic upward bias of $\widehat{\pi}_0$ in equation 26, which does not shrink with m and is bounded below by the impurity of Appendix C.8. The reference rule in Figure 2D uses the true π_0 and an f_1 estimated from independent data; the gap between it and the plug-in rule isolates the contribution of $\widehat{\pi}_0$. Direction matters: an error that lowers $\widehat{L}$ makes the rule conservative (smaller $\widehat{\mathcal{A}}_\tau$, FVR still $\leq 1 - \tau$); an error that raises $\widehat{L}$ makes it anti-conservative. Storey's bias is of the second kind.

C.8 Proposition 2: impurity, and what no p-value method can fix

The floor of the p-value histogram is what $\widehat{\pi}_0$ reads. If some invalid triplets produce p-values spread across the whole interval, they add to the floor, and the floor is higher than π_0. No estimator can subtract what the data cannot distinguish.

Proof of (i). Since $f_1 \geq \theta$ almost everywhere, $f_P \geq \pi_0 + (1-\pi_0)\theta = \pi_0^\star$ almost everywhere; and for every $\varepsilon > 0$ the set $\{f_1 < \theta + \varepsilon\}$ has positive measure, on which $f_P < \pi_0^\star + (1-\pi_0)\varepsilon$,

so $\operatorname{ess\,inf} f_P = \pi_0^\star$. Any two-group representation $f_P = \pi \cdot 1 + (1-\pi)g$ with $g \geq 0$ a density has $\pi \leq \operatorname{ess\,inf} f_P = \pi_0^\star$, and $\pi_0^\star$ is attained by $g^\star = (f_1 - \theta)/(1-\theta)$, which is nonnegative and integrates to one. Thus $(\pi_0^\star, g^\star)$ and (π_0, f_1) generate the identical f_P; the data are draws from f_P, so every statistic has the same distribution under both, and π_0 is not identified, only $\pi_0^\star$. For Storey's estimator, $\bar{F}_1(\lambda) = \int_\lambda^1 f_1 \geq \theta(1-\lambda)$, so by equation 26 $\mathbb{E}\,\widehat{\pi}_0(\lambda) \geq \pi_0 + (1-\pi_0)\theta = \pi_0^\star$. □

Proof of (ii). The best identifiable score is $L^\star(p) = \pi_0^\star / f_P(p)$, and the true one is $L(p) = (\pi_0/\pi_0^\star)L^\star(p)$. On the acceptance region $\mathcal{R} = \{p : L^\star(p) \geq \tau\}$, Lemma 1 gives $1 - \mathrm{FVR} = \mathbb{E}[L(P) \mid P \in \mathcal{R}] = (\pi_0/\pi_0^\star)\mathbb{E}[L^\star(P) \mid P \in \mathcal{R}]$, and $\tau \leq L^\star \leq 1$ on $\mathcal{R}$, so $1 - \pi_0/\pi_0^\star \leq \mathrm{FVR} \leq 1 - \tau\pi_0/\pi_0^\star$. □

Part (iii): noncentral alternatives. In addition to (i) and (ii): if invalid statistics are noncentral $\chi^2_q(\nu_i)$, then f_1 is decreasing and $\theta = f_1(1) = \mathbb{E}[e^{-\nu_i/2} \mid H_i = 1]$; in particular L is nondecreasing in p, which is what makes the rule $L(P_i) \geq \tau$ a p-value cutoff (Section 2.2).

Proof of (iii). The noncentral $\chi^2_q(\nu)$ density is $g_\nu(t) = \sum_{k\geq 0} e^{-\nu/2}\frac{(\nu/2)^k}{k!}\, g_{q+2k}(t)$. The ratio $g_{q+2k}(t)/g_q(t)$ is proportional to t^k, increasing in t, so g_ν/g_q is increasing in t (monotone likelihood ratio); since $t_p = F_q^{-1}(1-p)$ decreases in p, $f_1(p) = g_\nu(t_p)/g_q(t_p)$ decreases in p. As $p \to 1$, $t_p \to 0$, and every term with $k \geq 1$ vanishes relative to the $k = 0$ term, so $f_1(1) = e^{-\nu/2}$. With ν_i varying across invalid triplets, f_1 is the mixture over their law and $\theta = \mathbb{E}[e^{-\nu_i/2} \mid H_i = 1]$; each component is decreasing, so the mixture is. □

Worked case. $\nu = 2.05$ (Appendix B.6) gives $\theta = e^{-1.025} = 0.359$, $\pi_0^\star = 0.191 + 0.809 \times 0.359 = 0.481$, $\pi_0/\pi_0^\star = 0.397$ and $\mathrm{FVR} \in [0.603, 0.643]$ at $\tau = 0.9$. For $\nu = 11.4$, $\theta = 0.003$ and the interval is $[0.013, 0.112]$; for $\nu = 45.6$ it is $[0.000, 0.100]$: the guarantee is restored as soon as the weakest violations become detectable.

Why this is not an estimator problem. World 1 has $\pi_0 = 0.2$ and $f_1(p) = 1.8 - 1.6p$; world 2 has $\pi_0 = 0.36$ and $f_1(p) = 2 - 2p$. Both give $f_P(p) = 1.64 - 1.28p$. Any estimator sees the same f_P and, if unbiased for anything, is unbiased for $\pi_0^\star = 0.36$. In world 1 the rule $L^\star \geq 0.9$ accepts $p \geq 1.24/1.28 = 0.969$ and its true contamination is 0.474 rather than the nominal 0.10; the raw rule $P > 0.95$ has contamination 0.490 in the same world. The posterior rule is not the culprit: the p-values cannot support any rule with lower contamination without an assumption on θ, and a fixed cutoff is hit at least as hard.

What it means for the pipeline. (a) θ is governed by the least detectable invalid triplets and by faithfulness failures ($\nu_i = 0$), which add $\mathbb{P}(\nu_i = 0 \mid H_i = 1)$ to θ at every N (Appendix B.5). (b) In the simulations the overestimate of $\widehat{\pi}_0$ is confined to the cells where f_1 separates weakly from f_0 (Figure 2C: very-small regime, error 0.55 at $N = 500$ falling to 0.02 at $N = 5 \times 10^4$; small regime within 0.02 from $N = 2000$) and the FVR miss follows it (Figure 2D); in the moderate and large regimes the rule meets its target at every N. (c) In the near-null regime $\theta \approx 1$ and (ii) reduces to $\mathrm{FVR} \in [1 - \pi_0, 1 - \tau\pi_0] = [0.81, 0.83]$, the accept-all level every rule shows there (Figure 8D, accept-all line); that regime is a power failure, and by equation 6 it carries negligible bias. (d) A lower confidence bound on π_0 corrects sampling error, not impurity; the remedies are more N, a test with more power against weak violations, or an explicit assumption bounding θ.

C.9 Dependence and effective information

$L(p)$ is defined from marginal densities, so its population value does not depend on the dependence among candidates; what dependence changes is the realized null and hence the precision of any estimate of L (Efron, 2007). Here that precision is the precision of the mixture fit: m dependent p-values estimate π_0 as precisely as $m_{\mathrm{eff}} = m\mathrm{Var}(\widehat{\pi}_0 \mid \rho = 0)/\mathrm{Var}(\widehat{\pi}_0 \mid \rho)$ independent ones (Section 3.2). Figure 7A, B vary dependence alone through a block-equicorrelated Gaussian copula with blocks of 50, leaving f_0, f_1 and π_0 fixed; failed `qvalue` fits are excluded and their rate reported. Dependence leaves $\widehat{\pi}_0$ centred and the mean FVR near target but makes both more variable, so that at $\rho \geq 0.6$ a single nonempty accepted set misses the 10% target with probability 0.4–0.5 at every m,

against 0.06 at $m = 3000$ under independence (Figure 7B). Inside the SEM, the 7140 overlapping triplets of a 36-gene panel carry the information of about 220 gene-disjoint ones (Section 3.2). For discovery: the nominal candidate count overstates the information available to the mixture, and any single analysis should be read with the variance of Figure 7B in mind, not only the mean of Figure 2D.

C.10 THE BIAS IDENTITY: PROOF, SIGNED VERSION, FINITE-SAMPLE DECOMPOSITION

Why $b_i = 0$ for a valid triplet. Each ordered pair of disconnected negative controls satisfies the PCL conditions of Appendix A.2, the linear identity equation 12 then gives $\beta^{\mathrm{PCL}}(Z_{ia}, Z_{ib}) = \beta$, and the denominator is $\mathcal{D} = \sigma_{ZW}\mathrm{Var}(\varepsilon_T) \neq 0$ by equation 14. All six ordered-pair estimands equal β, so $\bar{\beta}_i = \beta$.

Proof of equation 6. $\sum_{i\in\mathcal{A}} |b_i| = \sum_{i\in\mathcal{A},\, H_i=0} |b_i| + \sum_{i\in\mathcal{A},\, H_i=1} |b_i| = 0 + I(\mathcal{A})\ \mathrm{Sev}(\mathcal{A})$; divide by $|\mathcal{A}|$.

Signed version and the pooled estimand. The quantity reported in practice is the pooled estimand $\bar{\beta}_{\mathcal{A}} = |\mathcal{A}|^{-1} \sum_{i\in\mathcal{A}} \bar{\beta}_i$, whose bias averages signed rather than absolute biases. Let $\mathrm{Sev}_{\pm}(\mathcal{A}) = I(\mathcal{A})^{-1} \sum_{i\in\mathcal{A},\, H_i=1} b_i$ be the mean signed bias of the accepted invalid triplets. The same split as in the proof of equation 6 gives

$$\bar{\beta}_{\mathcal{A}} - \beta = \frac{1}{|\mathcal{A}|} \sum_{i\in\mathcal{A}} b_i = \widehat{\mathrm{FVR}}(\mathcal{A})\ \mathrm{Sev}_{\pm}(\mathcal{A}), \tag{27}$$

and $|\,\mathrm{Sev}_{\pm}\,| \leq \mathrm{Sev}$, with equality if and only if all nonzero b_i share a sign; hence $|\bar{\beta}_{\mathcal{A}} - \beta| \leq \widehat{\mathrm{FVR}}\ \mathrm{Sev}$, the bound stated in Section 2.3. The identity for absolute biases is therefore an upper bound for the bias of the pooled estimate, tight exactly when there is no cancellation among the accepted invalid triplets. Cancellation cannot be relied on: it depends on the signs of the $V \to O$ paths of the accepted invalid triplets, which are unknown in data, and a benchmark in which invalid biases happened to cancel would credit the cancellation to the screening rule. The V-factor design therefore chooses λ so that all invalid biases share one sign (Appendix D.2); there the bound is an equality and the signed error in Figure 2F equals $\widehat{\mathrm{FVR}} \times \mathrm{Sev}$, and the sign-cancellation experiment of Appendix D.4 varies the signs deliberately to exhibit the gap between $|\,\mathrm{Sev}_{\pm}\,|$ and Sev.

Relevance in finite samples. The pooled designs keep every denominator away from zero, and in a sample $\widehat{\bar{\beta}}_i$ averages the ordered pairs whose $|\widehat{\mathcal{D}}|$ exceeds the numerical guard of Appendix A.4 (0.01 in the simulations, a relative tolerance in TCGA); it is undefined if no pair passes. This guards against the finite-sample blow-up of a ratio, not against invalidity: it does not change the estimand $\bar{\beta}_i$ of Section 2.3.

For discovery. The two factors are the two things one can and cannot buy with a better screening rule. $\widehat{\mathrm{FVR}}$ falls as the rule improves or as separation grows with N; Sev does not move. Figure 7C shows the product at most 0.015 for $\delta \geq 1$ against 0.29 at $(\delta, \lambda) = (0.25, 3)$, and the near-null regime is the other corner: $\widehat{\mathrm{FVR}} \approx 0.81$ but $\mathrm{Sev} \approx 10^{-10}$, so the accepted set is contaminated and harmless (Figure 8D and Figure 2F, near-null curves). A screening rule should be judged by the product, and the product by its factors.

Finite-sample decomposition. Write

$$\widehat{\bar{\beta}}_{\mathcal{A}} - \beta = \big(\widehat{\bar{\beta}}_{\mathcal{A}} - \bar{\beta}_{\mathcal{A}}\big) + \big(\bar{\beta}_{\mathcal{A}} - \beta\big). \tag{28}$$

The second term is the population structural bias, equal to $\widehat{\mathrm{FVR}}\ \mathrm{Sev}_{\pm}$ by equation 27. The first term has two parts. For a fixed set $\mathcal{A}$, each $\widehat{\beta}^{\mathrm{PCL}}$ is a smooth function of the sample covariance, so $\widehat{\bar{\beta}}_i - \bar{\beta}_i$ is $O_p(N^{-1/2})$ with an $O(1/N)$ plug-in bias (Appendix A.5). But $\mathcal{A}$ is not fixed: it is selected through the tetrad statistics $\widehat{\boldsymbol{t}}(\widehat{\Sigma})$, which share sample covariances with the PCL estimates, so the estimation error conditional on selection need not have mean zero. This is the post-selection

error. Its systematic part is small: on the (δ, λ) grid of Section 3.2 ($N = 5000$, moderate regime) the mean signed error of $\widehat{\bar{\beta}}_{\mathcal{A}}$ is within 0.013 of the signed structural term $\widehat{\mathrm{FVR}}\ \mathrm{Sev}_{\pm}$ in every cell, and on the pure-triplet grid (210 cells: six π_0, five regimes, seven N; $\tau = 0.9$) the median discrepancy is 0.001 with 95% of cells within 0.015; the largest departures (up to 0.048) occur at $N \leq 2000$ in the large regime and carry the sign of the ratio estimator's own finite-sample bias rather than that of the structural term, consistent with the $O(1/N)$ plug-in term dominating there. Re-estimating with the accepted set frozen and the population rebuilt at different $V \to O$ loadings reproduces equation 6 exactly (Appendix D.4). Its random part is not small: where the structural term exceeds 0.02 the RMSE of $\widehat{\bar{\beta}}_{\mathcal{A}}$ is 1.1 to 1.7 times larger on the (δ, λ) grid and up to 14.5 times larger on the pure-triplet grid (median 1.3), and where the structural term is near zero the RMSE sits on the sampling-variance floor (Figure 7D). Screening can make the structural term negligible; it cannot reduce sampling error.

D Simulation designs, results and figure guide

This appendix specifies every simulation of Section 3, records the results that the main text quotes but does not show, and states how each panel of Figures 2, 7 and 8 is computed. All numbers come from one Python package released with the paper (`vfactor_sim`), with $\widehat{\pi}_0$ and $\widehat{L}$ from the R package `qvalue` through a fixed interface (Appendix C.6); parameters are deterministic functions of the configuration and seeds are derived from configuration, regime, N and replicate index, so every cell reproduces draw for draw.

D.1 Conventions

(i) $\widehat{\mathrm{FVR}}_r$ is undefined for an empty accepted set; means of $\widehat{\mathrm{FVR}}$ are over replicates with $R_r > 0$ and the abstention probability $\mathbb{P}(R_r = 0)$ is reported alongside. (ii) A failed `qvalue` fit is recorded with $\widehat{\pi}_0$ missing and no accepted set; in the copula experiment such replicates (1771, all at high dependence and small m) are counted separately from abstention, and in the pure-triplet grids they enter as empty sets, which affects only $\pi_0 = 0$ (3390 of 8750 replicates) and three replicates at $\pi_0 = 0.10$. (iii) Coverage is undefined at $\pi_0 = 0$; elsewhere its mean includes abstaining replicates as zero. (iv) PCL quantities are conditional on a nonempty accepted set and are shown with the availability $\mathbb{P}(\mathcal{A} \neq \emptyset)$. (v) The headline threshold is $\tau = 0.90$ throughout the main text; the λ-field, sign-symmetric and mixed-panel runs were generated at $\tau = 0.70$ and are so labelled. (vi) Standard errors of Monte Carlo means are over independent replicate datasets, never over triplets within a dataset, because triplets within a dataset are dependent. No finite-sample bound for the plug-in rule is claimed; realized PCL error is measured against the known β on the same data used for selection, and no post-selection intervals are reported.

D.2 The V-factor benchmark

Model and parameters. Table 6 lists every coefficient of equation 7, its role, and whether it varies; Table 7 lists the five violation regimes and what each means for detectability and harm. Data follow equation 7 with $U, V \sim \mathcal{N}(0, 4)$ independent, all disturbances $\mathcal{N}(0, 1)$, $\beta = 0.5$, $a = c = 1.15$; u-loadings evenly spaced on $[0.8, 1.5]$ within each panel; violation loadings $v_i = \delta r_i u_i$ with $\delta = 1$ unless stated. The ratios r_i are evenly spaced over the regime range, near null $(10^{-7}, 5 \times 10^{-6})$, very small $(0.05, 0.15)$, small $(0.15, 0.40)$, moderate $(0.40, 0.90)$, large $(0.90, 1.80)$, in a permuted order such that $|r_i - r_j|$ is at least the grid step for every pair; this keeps every panel off the rank-one manifold $v_i/u_i \equiv$ const on which all tetrads vanish. The $V \to O$ loading is $\lambda = 1.5\, c\, \delta \max_i r_i$ unless set directly (8.4×10^{-6}, 0.25, 0.68, 1.53, 3.07 from near null to large). A second factor is used rather than direct gene-to-O edges because each such edge would add its variance to σ_{OO} and deflate every O-anchored statistic as the panel grows, whereas V adds a fixed $\lambda^2 \mathrm{Var}(V)$.

The regime is a statement about the detectability of invalid triplets, not about their harm. Harm is set by λ; the two are deliberately tied in the benchmark, where λ grows with the regime, and deliberately untied on the (δ, λ) surface, which is why that surface exists.

Closed forms. With $\kappa := \beta a + c$, the population covariances are $\sigma_{TT} = 4a^2 + 1$, $\sigma_{Tg_i} = 4au_i$, $\sigma_{g_i g_j} = 4(u_i u_j + v_i v_j)$, $\sigma_{g_i O} = 4\kappa u_i + 4\lambda v_i$, $\sigma_{TO} = \beta \sigma_{TT} + 4ac$, with $v_j = 0$ for valid-panel

Table 6: Coefficients of the V-factor benchmark equation 7: role, value, and whether it varies across cells.

symbol	role	value
u_i	proxy strength: loading of gene i on the confounder U	fixed; evenly spaced on $[0.8, 1.5]$ within each panel; identical in every cell
r_i	relative V-loading, v_i/u_i: how strongly gene i reflects V relative to U	range set by the regime (Table 7); evenly spaced within it, no two genes equal
v_i	loading on the second factor V; $v_i \neq 0$ makes a gene invalid	$v_i = \delta r_i u_i$; $\delta = 1$ everywhere except the (δ, λ) surface
λ	$V \to O$ loading: harm of an invalid triplet; also enters the O-anchored tetrads	$1.5\, c\, \delta \max_i r_i$ in the benchmark (8.4×10^{-6}, 0.25, 0.68, 1.53, 3.07 from near null to large); set directly on the surface
a, c	confounding paths $U \to T$, $U \to O$	fixed at 1.15, 1.15
β	target effect $T \to O$	fixed at 0.5
$\mathrm{Var}(U)$, $\mathrm{Var}(V)$	factor variances	fixed at 4, 4; all disturbances $\mathcal{N}(0, 1)$
π_0	proportion of valid triplets among the tested candidates	set by the panel sizes (k_0, k_1) (Table 8)
N	sample size	500 to 5×10^4, seven values
ρ	latent block correlation	copula experiment only; no SEM counterpart

Table 7: The five violation regimes. ν is the noncentrality of the joint test for a representative invalid triplet at $N = 2000$ (Appendix B.6); N^* the smallest N studied at which the mean error of $\widehat{\pi}_0$ is within 0.02; median $|b_i|$ the population PCL bias of invalid triplets.

regime	range of r_i	λ	ν at $N = 2000$	N^*	meaning
near null	$(10^{-7}, 5 \times 10^{-6})$	8.4×10^{-6}	5.7×10^{-18}	never	indistinguishable from valid at any N; median $\lvert b_i \rvert \approx 10^{-10}$, so harmless: a power failure, not a blind spot
very small	$(0.05, 0.15)$	0.25	2.05	5×10^4	sub-detection at practical N; median $\lvert b_i \rvert$ 0.04–0.06, below the severity gate
small	$(0.15, 0.40)$	0.68	45.6	2000	detectable from $N \approx 2000$; passes the severity gate
moderate	$(0.40, 0.90)$	1.53	167	500	detected at every N studied
large	$(0.90, 1.80)$	3.07	236	500	detected at every N studied; most harmful

genes. With $M_{xy} := v_x u_y - u_x v_y$,

$$T_{XY|ZT} = 16a\, v_z M_{xy}, \qquad T_{XY|ZO} = 16\, M_{xy}(\kappa v_z - \lambda u_z), \tag{29}$$

and for an ordered pair (Z, W) of panel genes

$$\beta^{\mathrm{PCL}}(Z, W) - \beta = \frac{4a\, v_Z\, (c\, v_W - \lambda u_W)}{u_Z u_W + \sigma_{TT} v_Z v_W}, \qquad \mathcal{D}(Z, W) = 4\big(u_Z u_W + \sigma_{TT} v_Z v_W\big), \tag{30}$$

both verified numerically to 10^{-14}. Four facts follow. A valid pair has $v_Z = v_W = 0$, so every tetrad vanishes, $\beta^{\mathrm{PCL}} = \beta$ and $\mathcal{D} = 4u_Z u_W \geq 2.56$. A tetrad needs two violation genes in its first pair to be nonzero, and with $v_i = \delta r_i u_i$, $M_{xy} = \delta u_x u_y (r_x - r_y)$, so every tetrad is proportional to a difference of ratios: this is why the ratios are kept apart, and why the T-anchored tetrads scale with δ^2 (detectability) while the O-anchored ones also depend on λ (severity and detectability are not orthogonal). Under the default λ the factor $c\, v_W - \lambda u_W = c\, u_W \delta(r_W - 1.5 \max r)$ is negative for every W, so every pair bias is negative and all b_i share one sign: the *aligned* design, in which signed and absolute accepted-set error coincide (the gate records the fraction of negative b_i, which is 1.00 in every aligned regime). And on the rank-one manifold $v_i = \kappa' u_i$ all tetrads vanish while the bias is $4a\kappa'(c\kappa' - \lambda)/(1 + \sigma_{TT}\kappa'^2) \neq 0$ in general, which is the blind spot of Appendix B.5; the near-null regime is different, since there $v_i \approx 0$ makes both tetrads and biases vanish. (Because $\kappa = 1.5c$ under the defaults, the O-anchored tetrads whose third slot is the gene with the largest ratio are identically

zero; the joint statistic is unaffected, and it explains why one of the six Wishart p-values is null-like for triplets containing that gene.)

Population gate. Before any Monte Carlo run the population covariance of a configuration is evaluated analytically and, for every triplet and all six ordered pairs, tetrads, b_i and denominators are recorded. Structural checks: valid tetrads and valid b_i zero to 1.4×10^{-14}; alignment margins $\lambda u_i - c v_i > 0$; ratio grid off the rank-one manifold. Downstream checks, enforced when a run supports severity claims: median invalid $|b_i| \geq 0.10$, tenth percentile ≥ 0.05, minimum $|\mathcal{D}| \geq 0.05$ over all ordered pairs. The very-small regime fails the severity checks by construction (median invalid $|b_i|$ of 0.04–0.06) and supports acceptance and contamination claims only; in the (δ, λ) surface 15 of 30 cells fail the same-sign condition and are marked rather than dropped.

Pure-triplet configurations and rules. A panel of k_0 valid and k_1 violation genes yields $m_0 = \binom{k_0}{3}$ valid and $m_1 = \binom{k_1}{3}$ invalid triplets, $\pi_0 = m_0/(m_0 + m_1)$ (Table 8); π_0 is a proportion of tested triplets, not of genes. $N \in \{500, 1000, 2000, 5000, 10^4, 2 \times 10^4, 5 \times 10^4\}$, 250 replicates per cell. Every rule is evaluated on the same fitted mixture within a replicate: accept-all; plug-in local validity at $\tau \in \{0.05, \ldots, 1\}$; the reference oracle where available; raw-p cutoffs $P > c$ on a 6-point grid (13-point in the transport and mixed runs); DANCE all-pass; Bonferroni ($P \geq 0.05/m$); and Wishart-component (Appendix B.7).

Table 8: Pure-triplet presets: $\pi_0 = \binom{k_0}{3}/\{\binom{k_0}{3} + \binom{k_1}{3}\}$, m the number of tested triplets.

violation panel k_1	valid panel k_0	valid	invalid	m	π_0
28	0	0	3276	3276	0.000
28	14	364	3276	3640	0.100
22	14	364	1540	1904	0.191
18	14	364	816	1180	0.308
18	16	560	816	1376	0.407
16	16	560	560	1120	0.500

Mixed panels. The mixed configurations enumerate all triplets of a panel and label each by its number K of violation genes, $H_i = 0 \iff K = 0$. From equation 29, a $K = 1$ triplet has four null tetrads of six (all three T-anchored and one O-anchored), $K = 2$ has one, $K = 3$ none; and a pair with a valid W has bias $-4a\lambda v_Z/u_Z$, which does not saturate in δ, so the mixed classes $K = 1, 2$ carry both the largest tetrad signal and the largest severity while the all-violation class $K = 3$, whose members all lie near the rank-one manifold, has the smallest of both. Runs: $16+16$ and $6+26$ panels (4960 triplets) and a 36-gene enumeration with $k_0 \in \{0, 17, 21, 25, 27, 29\}$ (7140 triplets; $\pi_0 = 0$, 0.095, 0.186, 0.322, 0.410, 0.512), regimes very small, small, moderate, 250 replicates, $\tau = 0.7$. Findings: $K = 1, 2$ are accepted at rates below 10^{-4}; only $K = 3$ leaks, at 0.109 for $N = 1000$ down to 0.0002 for $N = 2 \times 10^4$ ($k_0 = 25$, small regime), so the accepted set's contamination is at most 1.3% and its harm contribution is an order of magnitude below the finite-sample PCL error. The all-violation pure design of the main text is therefore the hard case for contamination and the easy case for severity.

D.3 Experiment constructions

Calibration and recovery (Figure 2B, C; Figure 8A–C). Null size: per dataset, the fraction of valid-triplet p-values below 0.05, averaged over 250×5 datasets with a 95% interval over datasets. QQ: valid-triplet p-values pooled across the nonzero-π_0 presets, all regimes and 20 replicates at $N \in \{500, 2000, 10^4\}$. Densities: $\pi_0 = 0.191$, $N = 5000$, f_0 pooled over replicates and regimes, f_1 per regime, 40-bin histograms on a log scale. Recovery: one `qvalue` fit per replicate; the plotted value is the mean of $\widehat{\pi}_{0,r} - \pi_0$ over the 250 replicates of a cell, the Wishart-component fit using the six component p-values per triplet.

Reference oracle against plug-in (Figure 2D). $\pi_0 = 0.191$, $N \in \{1000, 2000, 5000, 10^4\}$, five regimes, 250 evaluation replicates per cell. Plug-in: one `qvalue` fit to the 1904 p-values of the evaluation dataset. Oracle: 100 disjoint-seed reference datasets per cell (154,000 invalid p-values)

give a 200-bin histogram $\widehat{f}_1$; $L_{\mathrm{ref}}(p) = \pi_0/\{\pi_0 + (1-\pi_0)\widehat{f}_1(p)\}$ with the exact π_0, made monotone by a cumulative maximum; the rule is $P_i \geq c^*(\tau)$ with $c^*(\tau)$ the left edge of the first bin with $L_{\mathrm{ref}} \geq \tau$. It is a high-precision approximation, not an analytic oracle, which is why a few weak-cell points lie slightly above the population bound. A reliability diagram of the plug-in score (30 replicates, small regime) shows scores near 0.99 corresponding to valid fractions of 0.69 at $N = 1000$ and 0.94 at $N = 10^4$.

Threshold transport (Figure 2E). One bank of 40 valid and 40 violation genes (9880 pure candidates of each class). Per simulated dataset, six lists of $m = 2000$ are sampled without replacement with $(m_0, m_1) \in \{(0, 2000), (200, 1800), (382, 1618), (616, 1384), (814, 1186), (1000, 1000)\}$; the lists share subjects, genes and covariance, so f_0 and f_1 are common to the six cells and only the mixture weight changes. Five regimes, seven N, 250 paired datasets. $c^*_{0.10}$ is obtained by bracketing 0.10 between adjacent recorded cutoffs of $\overline{\mathrm{FVR}}(c) = \mathbb{E}[\widehat{\mathrm{FVR}}_r(c) \mid R_r(c) > 0]$ and interpolating in $\log c$; an up-triangle means $\overline{\mathrm{FVR}}(0.5) > 0.10$, a down-triangle $\overline{\mathrm{FVR}}(0.001) \leq 0.10$. The global cutoff maximizes calibration-split coverage (replicates 0–124) subject to average contamination ≤ 0.10 over the identifiable cells (small, moderate, large; $\pi_0 > 0$): the candidates $c \in \{0.05, 0.075, 0.10\}$ give contamination 0.101, 0.093, 0.087 and coverage 0.953, 0.928, 0.903, so $c = 0.075$; it is evaluated on replicates 125–249. Including the very-small regime, the minimum attainable average contamination is 0.154 (all regimes 0.262), both at $c = 0.5$.

Dependence (Figure 7A, B). Tests are divided into blocks of 50; $Z_i = \sqrt{\rho}\, B_{b(i)} + \sqrt{1-\rho}\, \varepsilon_i$ with $B_b, \varepsilon_i \sim \mathcal{N}(0,1)$, and $U_i = \Phi(Z_i)$, so every marginal is uniform whatever ρ and two tests in a block have latent correlation ρ. With $\pi_0 = 0.191$, valid tests take $P_i = U_i$ and invalid tests $P_i = F_1^{-1}(U_i)$ with F_1 the empirical distribution of small-regime invalid Bollen p-values; $\rho \in \{0, 0.1, 0.3, 0.6, 0.9\}$, $m \in \{250, 500, 1000, 2000, 3000\}$, 1000 replicates; sensitivity arms with block size 10 and with the moderate F_1 give the same ordering by ρ. Structural control inside the SEM: in the 36-gene panel with $k_0 = 21$ ($\pi_0 = 0.186$), 220 gene-disjoint triplets, 220 overlapping triplets with the same K census, and the full 7140; sharing genes raises the SD of $\widehat{\pi}_0$ by about 35% at $m = 220$, and the full enumeration only brings it back to that of 220 disjoint triplets.

Frozen-selection identity check. The accepted sets at $\tau \in \{0.5, 0.7, 0.9\}$ of one screening replicate (small regime, $N = 2000$, $\pi_0 = 0.191$; $|\mathcal{A}| = 436, 384, 308$ with $97, 64, 39$ invalid, $\widehat{\mathrm{FVR}} = 0.222, 0.167, 0.127$) are frozen; the population is rebuilt with $\lambda \in \{0, 0.2, \ldots, 2\}$ and b_i recomputed for the frozen members. $\mathrm{Mean}_{\mathcal{A}}|b_i|$ against $\mathrm{Sev}(\mathcal{A})$ lies on a ray through the origin with slope $\widehat{\mathrm{FVR}}(\mathcal{A})$; over a global pool of 800 accepted sets (regimes small and moderate, three N, both statistics, five λ), the 670 with $I(\mathcal{A}) > 0$ lie on $y = x$ with maximum residual 4.4×10^{-16}. This is a numerical check of the identity equation 6, not evidence about the method.

Detectability and severity (Figure 7C, D). (δ, λ) surface: $\delta \in \{0.25, 0.5, 1, 1.5, 2\}$, $\lambda \in \{0.25, 0.5, 1, 1.5, 2, 3\}$, moderate regime, $\pi_0 = 0.191$, $N = 5000$, 250 replicates, full pipeline rerun per cell. Sign-symmetric arm: $\mathrm{sign}(v_i)$ alternated over the violation panel with $\lambda > 0$ unchanged, so the b_i are not symmetric about zero (fraction negative 0.76, 0.11, 0.13 in the small, moderate and large regimes) and cancellation is partial; it exhibits the gap between $|\,\mathrm{Sev}_{\pm}\,|$ and Sev of Appendix C.10. A λ sweep at fixed δ shows contamination falling and severity rising with λ while their product moves much less than either factor.

Finite-sample PCL evaluation. For each accepted triplet the six ordered assignments are evaluated with equation 1 on the replicate's sample covariance; a pair is discarded when $|\widehat{\mathcal{D}}| < 0.01$; the triplet estimate is the mean of the retained pairs and $\widehat{\widehat{\beta}}_{\mathcal{A}}$ the equal-weight mean of triplet estimates. Recorded: signed, absolute and squared error against $\beta = 0.5$, conditional on $\mathcal{A} \neq \emptyset$.

Estimation-budget control. On the same datasets ($\pi_0 = 0.191$, small and moderate, five N): Bollen ($m = 1904$); W1, one uniformly chosen tetrad per triplet as fit and test; W6-matched-m, the W1 fit applied to all six tetrads with the all-six rule; W6-full, fit on all 11,424 tetrad p-values; and Bollen pooled over six independent datasets. Every Wishart fit is biased upward by the same amount whatever its count (W6-full and W1 agree to 0.005), so the bias is a property of the per-tetrad statistic;

count buys precision only, and less for Wishart ($\mathrm{Var}(\mathrm{W1})/\mathrm{Var}(\text{W6-full}) = 2.2$–$3.3$) than for Bollen with independent data (4.6–7.8); W6-full and W6-matched-m coincide downstream, so the Wishart family's low contamination and low coverage come from the all-six rule, not from the count.

D.4 Results quoted in the main text

At $\tau = 0.9$ and $\pi_0 = 0.191$: plug-in contamination 0.76 at $N = 1000$ and 0.44 at $N = 10^4$ in the very-small regime; $0.31/0.20/0.10/0.04$ in the small regime at $N = 1000/2000/5000/10^4$, with abstention $0.52/0.44/0.46/0.04$; $0.10/0.04/0.01/0.002$ in the moderate regime; the oracle abstains in every replicate of the small regime at $N = 1000$ and has $0.12/0.06/0.03$ at the other three N (Figure 2D). Accept-all fixes contamination at 0.809; DANCE all-pass has 0.73 at $N = 1000$ falling to 0.33 at $N = 10^4$ in the small regime, and the Wishart-component rule reaches a lower contamination than Bollen at large N only by keeping a far smaller share of the valid triplets (Figure 8D, E). The plug-in rule retains 55–60% of valid triplets in the moderate and large regimes (Figure 8E). In the V-factor benchmark the replicate standard deviation of $\widehat{\pi}_0$ stays near 0.10 at every N. On the pure-triplet grid (210 cells, $\tau = 0.9$) the realized signed error of $\widehat{\bar{\beta}}_{\mathcal{A}}$ tracks the signed structural term with median discrepancy 0.001, 95% of cells within 0.015 and a maximum of 0.048 at $N \le 2000$ in the large regime, where the finite-sample bias of the ratio estimator dominates and has the opposite sign; where the structural term exceeds 0.02 the RMSE is 1.0–14.5 times larger (median 1.3). On the (δ, λ) surface contamination falls from 0.49 at $(0.25, 0.25)$ to below 0.002 at $\delta = 2$, severity rises with λ and saturates in δ ($\mathrm{Sev} \approx 0.56$ at $(2, 3)$), the product is 0.29 at $(0.25, 3)$ against at most 0.015 for $\delta \ge 1$, the RMSE runs from 0.02 to 0.32, the mean signed error is within 0.013 of the signed structural term in every cell, and where that term exceeds 0.02 the RMSE is 1.1–1.7 times larger (Figure 7C, D).

D.5 Figure guide

Figure 2. A: the design of equation 7. B: 40-bin log-density histograms at $\pi_0 = 0.191$, $N = 5000$, f_0 pooled, one f_1 per regime. C: mean of $\widehat{\pi}_{0,r} - \pi_0$ over replicates, one line per regime, band over the six π_0 presets. D: $\mathbb{E}[\widehat{\mathrm{FVR}}_r \mid R_r > 0] - 0.10$ at $\tau = 0.9$, plug-in (filled) and reference oracle (open), $\pi_0 = 0.191$. E: $c^*_{0.10}$ per cell as defined above, one facet per $\pi_0 \in \{0.1, 0.191, 0.5\}$, tuned $c = 0.075$ dashed. F: mean of $\widehat{\bar{\beta}}_{\mathcal{A}} - 0.5$ over nonempty replicates at $\tau = 0.9$; dotted, the same without screening.
Figure 8. A: the densities of Figure 2B, one facet per regime. B: ordered valid-triplet p-values against $(j - \frac{1}{2})/M$, pooled as above, at $N = 500$ and 10^4. C: as Figure 2C with the Wishart-component fit dashed. D: $\mathbb{E}[\widehat{\mathrm{FVR}}_r \mid R_r > 0]$ at $\tau = 0.9$ on a log scale for Bollen plug-in (solid), Wishart-component (dashed), DANCE all-pass (dash-dot) and accept-all (dotted). E: mean coverage over all replicates, abstentions as zero, same rules. F: as Figure 2F with the comparator rules. G: $\mathbb{P}(R_r = 0)$ for Bollen plug-in and Wishart-component.
Figure 7. A: m_{eff}/m as defined in Section 3.2, SDs over successful fits, one line per m. B: $\#\{r : R_r > 0, \widehat{\mathrm{FVR}}_r > 0.10\}/\#\{r : R_r > 0\}$ at $\tau = 0.9$, one line per ρ. C: cell mean of $\widehat{\mathrm{FVR}}(\mathcal{A}_r)\,\mathrm{Sev}(\mathcal{A}_r)$ over nonempty replicates. D: RMSE of $\widehat{\bar{\beta}}_{\mathcal{A}}$ over nonempty replicates; $\times$ marks the 15 gate-failed cells.

E Acceptance-side terminology, fixed rules, comparison and modularity

E.1 Exact relationship to false nondiscovery

Let m hypotheses be tested, W the number not rejected and T the number of alternatives among them. The false nondiscovery rate of Genovese & Wasserman (2002) is $\mathrm{FNR} = \mathbb{E}[T/\max(W, 1)]$ and its positive version (Storey, 2003) conditions on $W > 0$; Sarkar (2006) gives FNR results for single-step procedures under dependence. Under the two-group model equation 3 with rejection region $\{p \le \gamma\}$, the pFNR of the nonrejection set $\{p > \gamma\}$ is $\mathbb{P}(H = 1 \mid P > \gamma)$, which is $\mathrm{FVR}(\gamma)$ of equation 4. Our accepted set is not the whole nonrejection set: we compute $L(p)$ and accept $\{i : L(P_i) \ge \tau\}$, so $\mathrm{FVR}(\mathcal{A}_\tau)$ is a pFNR evaluated on the classifier's acceptance region. The name FVR is domain-specific, chosen because an accepted null is passed downstream as a causal proxy; generic FNR and pFNR theory, the two-group model and Storey's estimators are

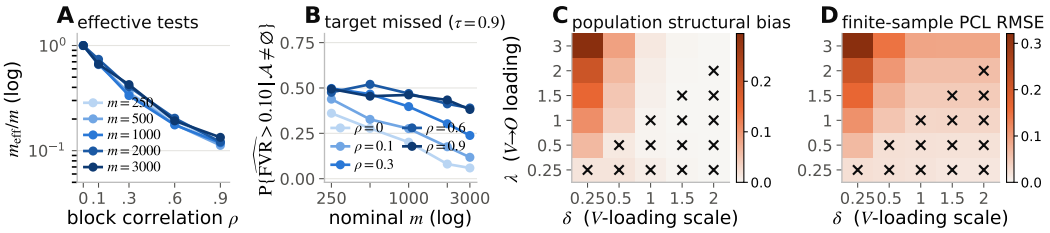


Figure 7: **Dependence between tests, and downstream bias. A, B:** p-value-level experiment with no SEM. m, $\pi_0 = 0.191$, f_0 and f_1 (the small-regime invalid Bollen p-values) are identical for all $\rho \in \{0, 0.1, 0.3, 0.6, 0.9\}$. Dependence enters only through a block-equicorrelated Gaussian copula, $z_i = \sqrt{\rho}\, B_{b(i)} + \sqrt{1-\rho}\, \varepsilon_i$ with blocks of 50, which leaves every marginal unchanged. 1000 replicates per cell; failed `qvalue` fits excluded (Appendix C.9). **A:** m_{eff}/m against ρ, one line per m. **B:** $\mathbb{P}(\widehat{\text{FVR}}_r > 0.10 \mid R_r > 0)$ at $\tau = 0.9$ against m, one line per ρ. **C, D:** full V-factor pipeline rerun at $\tau = 0.9$ in each of 30 worlds (moderate regime, $\pi_0 = 0.191$, $N = 5000$, 250 datasets per cell), varying the violation scale δ ($v_i = \delta r_i u_i$) and the $V \to O$ loading λ. **C:** mean $\widehat{\text{FVR}} \times \text{Sev}$, the accepted set's mean absolute structural bias. **D:** RMSE of $\widehat{\beta}_{\mathcal{A}}$. C is pale at large δ because contamination collapses ($\widehat{\text{FVR}}$ falls from 0.49 at $\delta = 0.25$ to 0.0015 at $\delta = 2$), not because the accepted invalid triplets are harmless (Sev ≈ 0.56 at $(\delta, \lambda) = (2, 3)$). $\times$: the 15 cells with $\lambda \leq c\,\delta \max_i r_i$, outside the design's same-sign condition; their biases are still one-signed in all but two.

prior work, and Proposition 1 is the posterior consequence of thresholding L, not a new theorem about multiple testing. Three quantities must not be confused: $\text{FVR}(\mathcal{A})$ is a population conditional probability; $\widehat{\text{FVR}}(\mathcal{A}) = I(\mathcal{A})/|\mathcal{A}|$ is its empirical analogue in one dataset, undefined when $|\mathcal{A}| = 0$; and valid-triplet coverage $\mathbb{P}(i \in \mathcal{A} \mid H_i = 0)$ measures how much of the valid class is retained.

E.2 Fixed-threshold and Bonferroni-type acceptance rules

DNCT validation is an acceptance problem, whereas classical multiplicity corrections are designed for rejections; none of the following fixed rules has a stable validation interpretation. *Fixed significance threshold.* $\mathcal{A}_\gamma = \{i : P_i > \gamma\}$; DANCE uses $\gamma = 1/N$ (Kummerfeld et al., 2024), a conventional choice is 0.05. Its FVR equation 4 depends on π_0 and f_1 and does not adapt to them; as N grows, $\gamma = 1/N$ shrinks, but whether invalid candidates pass depends on how fast their p-values shrink as well, so the contamination is again a property of the mixture, not of the number $1/N$ (Figure 2E; Appendix C.5). *Bonferroni, rejection side.* Rejecting when $P_i < \alpha/m$ bounds the family-wise error of rejections; since we accept what is not rejected, the acceptance set is $\{P_i \geq \alpha/m\}$, with $m \approx 10^3$ and $\alpha = 0.05$ a cutoff of 5×10^{-5}, far below most weak-alternative p-values, so nearly all triplets are accepted, invalid ones included. Controlling the rejection-side error translates into being maximally permissive about acceptances; Benjamini–Hochberg (Benjamini & Hochberg, 1995) targets rejections in the same way. *Bonferroni for acceptances.* Bounding $\mathbb{P}(\geq 1 \text{ false acceptance}) \leq \alpha_A$ would require per-test power at least $1 - \alpha_A/m_1$ against every alternative, with m_1 the unknown number of invalid triplets; with $m_1 \approx 800$ and $\alpha_A = 0.05$ this is power 0.99994, attainable only for strong violations or very large N, and uniform over alternatives regardless of effect size. *Local validity.* $\widehat{\mathcal{A}}_\tau = \{i : \widehat{L}(P_i) \geq \tau\}$ has the oracle property of Proposition 1 adaptively: its effective cutoff on p adjusts to $\widehat{\pi}_0$ and to the shape of f_1. In the simulations DANCE all-pass and rejection-side Bonferroni are the most permissive rules in every cell (Figure 8D).

Multiplicity in applied tetrad analysis. Applied vanishing-tetrad analyses have rarely addressed the multiplicity of the constraints implied by a measurement model. Single-tetrad tests (Wishart, 1928) were long reported one constraint at a time. Bollen (1990) and Bollen & Ting (1993) instead test the nonredundant tetrads implied by a measurement model simultaneously with one chi-square statistic, and Gudergan et al. (2008), who test the tetrads separately, recognize that this inflates the type I error and use Bonferroni-adjusted confidence intervals. These corrections address false rejections across constraints; proxy validation reverses the consequential direction, and a corrected

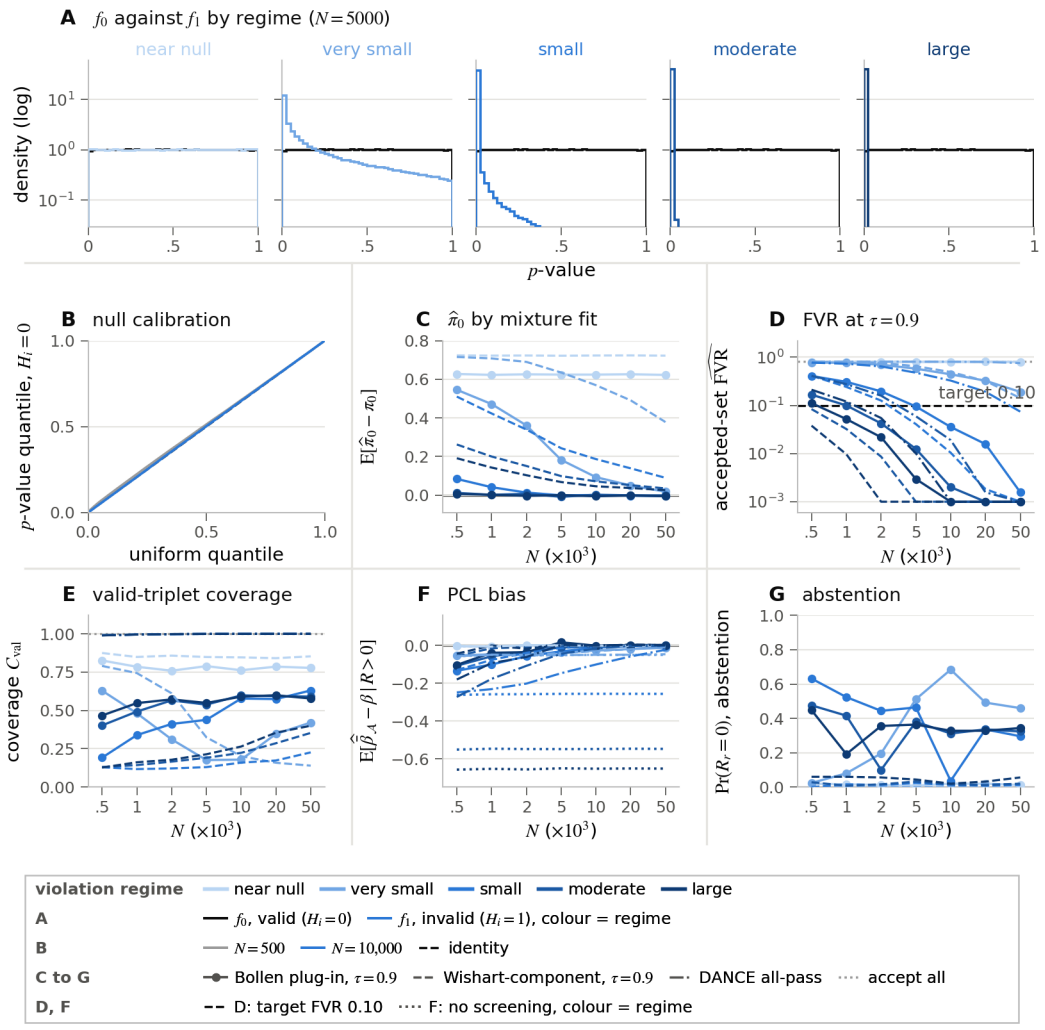


Figure 8: **Comparators and diagnostics for the V-factor benchmark** ($\pi_0 = 0.191$ unless stated; regime colours as in Figure 2). **A:** p-value densities of valid (f_0, black) and invalid (f_1, colour) triplets by regime at $N = 5000$. **B:** null calibration: quantiles of the valid-triplet p-values against the uniform, $N = 500$ and $N = 10^4$. **C:** mean signed error of $\widehat{\pi}_0$; solid Bollen, dashed Wishart-component (mixture fitted to the $6m$ per-tetrad p-values). **D:** accepted-set FVR at $\tau = 0.9$ over nonempty replicates; solid Bollen plug-in, dashed Wishart-component, dash-dot DANCE all-pass, dotted accept-all; the horizontal line is the 0.10 target. **E:** valid-triplet coverage over all replicates (abstentions counted as zero). **F:** signed finite-sample bias of $\widehat{\widehat{\beta}}_{\mathcal{A}}$ over nonempty sets; dotted, no screening. **G:** abstention probability $\mathbb{P}(R_r = 0)$.

rejection-side procedure accepts everything it does not reject. We are not aware of applied tetrad work that targets the acceptance-side error, but we have not conducted a systematic review.

E.3 Comparison with proxy-discovery methods

Ours is the only row whose decision rule targets the composition of the accepted set, and the only one that links the acceptance error to the estimator's structural bias equation 6. For the generalized tetrad test, the standardized statistic controls the type I error for levels $\alpha \leq 0.215$, and the two generalized tetrad differences of a quadruple are combined into a single aggregated statistic rather than tested separately (Ying et al., 2025); both concern the size of a rejection-side test.

Table 9: Proxy-discovery methods by structural signal, main assumptions, unit tested, acceptance rule, whether erroneous acceptance across the candidate search is the controlled error, and downstream causal use. Cells marked † follow the published descriptions and were not re-verified against the source.

Method	Structural signal	Main assumptions	Unit tested	Acceptance rule	Acceptance-side control	Downstream causal use
DANCE (Kummerfeld et al., 2024)	covariance tetrads (six per triplet)	linear SEM, single latent confounder	each tetrad; triplet by all-pass	no Wishart test rejects at α, default $\alpha = 1/N$	none targeted; α is a heuristic	yes: NC estimation on validated DNCTs
Proxy-Rank (Xie et al., 2024)	rank constraints (second-order statistics)	linear model; latent dimension specified†	candidate proxy sets	fixed-level rank tests†	not targeted†	yes: causal-effect estimation
Proxy-GIN (Xie et al., 2024)	non-Gaussian independence (higher-order statistics)	linear non-Gaussian model	candidate proxy sets	independence tests†	not targeted†	yes: causal-effect estimation
Generalized tetrads (Ying et al., 2025)	bridge-function tetrad constraint	nonlinear or non-parametric latent model	one aggregated statistic over the two generalized tetrad differences of a quadruple	calibrated test of the constraint ($\alpha \leq 0.215$)	rejection-side, not proxy-set FVR	not the paper's object; usable as the test in a discovery pipeline
Ours	Bollen joint tetrads (any calibrated candidate test)	primary analysis linear, effective one-factor	triplet	$\widehat{L}(P) \geq \tau$	acceptance-side FVR through local validity; oracle bound, quantified plug-in error	yes: triplet-averaged PCL with FVR$\times$severity accounting

Extended related work. DANCE is our closest starting point. Under its linear single-confounder model a DNCT behaves as three measurements of a common latent factor and satisfies vanishing tetrad constraints (Spearman, 1904); DANCE evaluates six tetrads separately with Wishart's determinant test (Wishart, 1928) and retains a triplet when none rejects at a threshold such as $1/N$. Our work changes two parts of this pipeline: the statistical object validated is the triplet, whose tetrads are correlated constraints on that object, so we use Bollen's joint covariance-aware statistic (Bollen, 1990) to obtain one candidate-level p-value; and once many overlapping triplets are examined we treat their validation as a multiple-testing and classification problem and target errors among the triplets that are retained. Proxy-Rank, Proxy-GIN and generalized tetrad procedures enlarge the class of latent structures that can be tested; our focus is complementary, since the generality of the structural test and the large-scale validation rule are separate questions. Given a calibrated candidate-level test, we ask how a large and dependent collection of candidate proxy sets should be converted into an accepted set for downstream causal estimation. To our knowledge no data-driven proxy-discovery method has treated this as an acceptance-side multiple-testing problem or linked validation error to the bias of the downstream estimate.

E.4 Modularity with respect to the structural test

The acceptance-side framework does not depend on Bollen's linear test. Let P_i be any candidate-level p-value that is calibrated under the structural null, $P_i \mid H_i = 0 \sim \mathrm{Unif}(0, 1)$, with alternatives stochastically smaller. The two-group mixture equation 3, the local validity equation 5, the acceptance set $\{i : \widehat{L}(P_i) \geq \tau\}$, the oracle bound of Proposition 1 and the factorization equation 6 proceed unchanged; only the map from data to P_i is replaced. The generalized tetrad test of Ying et al. (2025) is the natural substitute when the latent relation is nonlinear, where Bollen's covariance tetrads are miscalibrated under the null and the mixture is misspecified. What the framework does require is calibration of the candidate-level test under the structural null, because $f_0 \equiv 1$ enters both $\widehat{\pi}_0$ and $\widehat{L}$; a miscalibrated test perturbs both jointly, which is why the simulations evaluate the fitted rule directly rather than reasoning from the sign of one error (Appendix B.9).

F TCGA–BRCA protocol and diagnostics

This appendix records the full-sample TCGA–BRCA run of Section 4 (`tcga/tcga_pipeline.py`). The released code reproduces the saved outputs of that run: test statistics to 2×10^{-11}, p-values to 10^{-13}, accepted sets and per-gene tables identical.

F.1 Data and panel

507 tumour samples with expression for 15,593 genes; the outcome is AJCC stage coded $0, 1, 2, 3$ and used as a number. The panel is the 700 genes with the smallest adjusted p-value in a `limma` analysis of stage, plus curated driver and housekeeping lists; every column is mean-centred; genes whose sample covariance with stage is below 0.05 in absolute value are dropped, housekeeping genes excepted. 611 genes remain; the released list pins them, because two genes tie at rank 700 and a different sort order admits a 612th.

F.2 Candidate treatments and proxy pools

Each of the 611 genes is in turn the treatment T. For that T the other 610 genes are clustered by complete-linkage hierarchical clustering on the distance $1 - |\widehat{\mathrm{corr}}|$, cut at 0.65, so that two genes share a cluster only if every pair within it has $|\widehat{\mathrm{corr}}| \geq 0.35$; the proxy pool of T is the cluster containing T, minus T. This is the relevance requirement of Appendix A.4 applied before testing. Treatments whose pool has fewer than four genes are skipped: 203 of 611 remain, with a median pool of 7 genes, and every triple of genes from a pool is a candidate triplet, $m = 5{,}710{,}820$ in all. Because pools overlap and every gene serves as a proxy for many treatments, the candidates are strongly dependent (Appendix C.9).

F.3 Test, fit, acceptance and estimation

For each candidate (A, B, C) with its T and stage as O, the four canonical tetrads are computed from the full-sample covariance, with the delta-method covariance, the ridge 10^{-7}, the pseudo-inverse Wald statistic and its χ^2_4 p-value, exactly the function of Appendix B.6; the numerical rank is 4 in every one of the 5.7 million tests. The six Wishart p-values of each triplet are also computed and their minimum stored for the DANCE comparison. One `qvalue` call on all m p-values (Appendix C.6) gives $\widehat{\pi}_0 = 0.157$ and $\widehat{L}$ per triplet. Accepting $\widehat{L} \geq \tau$ retains 557,861 triplets at $\tau = 0.5$, 220,927 at 0.7 and 45,270 at 0.9 (9.8%, 3.9% and 0.8% of the candidates); DANCE's rule, smallest Wishart p-value above $1/507$, retains 3,198,500 (56%). For each accepted triplet the six ordered-pair PCL estimates equation 1 on the full-sample covariance are averaged, a pair being dropped when its denominator is below 10^{-4} times $|\widehat{\sigma}_{TT}\widehat{\sigma}_{ZW}| + |\widehat{\sigma}_{TW}\widehat{\sigma}_{TZ}|$ (Appendix A.4). For each gene the pipeline records how many accepted triplets contain it as a proxy, how many have it as the treatment, and the mean and median of the triplet estimates over those.

F.4 How Figure 3 is computed

Panel A: the density of all m p-values in 40 equal bins (area one), axis cut at 1.2 so that the right tail is visible, with a dashed line at $\widehat{\pi}_0 = 0.157$. If the two-group model holds, the density far from zero is the null component alone and its height is π_0; the fit coincides with the flat right tail. Panel B: one point per gene, x the number of the 45,270 accepted triplets containing the gene as a proxy, y the median of the accepted-triplet PCL estimates in which the gene is the treatment; genes never the treatment of an accepted triplet are drawn as crosses at $y = 0$; the most selected proxies and the largest $|y|$ are labelled. The most selected proxies (CKS1B, SGOL1, GSG2, KPNA2, TRIP13, DONSON, DCC1, each in 3,000 to 3,800 accepted triplets) have medians within ± 0.03 of zero, and the largest estimates (0.22 for PLK1 and UBE2C, -0.19 for CDCA8) belong to genes selected fewer than 500 times. The separation is confined to the upper tail: over genes, the median $|\widehat{\widehat{\beta}}|$ is 0.040 both for the top decile of proxies and for genes selected at most five times (at $\tau = 0.7$), but the 90th percentiles are 0.09 against 0.14 and the fractions above 0.1 are 6% against 14%.

F.5 Six-estimate agreement diagnostic

Two consequences of the DNC definition are checkable without knowing β (Section 2.1). Any two DNCs of the same U identify the same estimand in either assignment, so the six ordered-pair estimates of an accepted triplet should agree; and a DNC has no causal effect on O, so its estimate in the treatment role should be near zero. Figure 9 reports the first for the TCGA candidates, comparing the 45,270 triplets accepted at $\widehat{L} \geq 0.9$ with 59,976 sampled triplets rejected at $\widehat{L} < 0.1$. Within a triplet, the standard deviation of the six estimates has median 0.017 among accepted and 0.062

among rejected triplets (panel A); the swap asymmetry $|\widehat{\beta}(Z,W) - \widehat{\beta}(W,Z)|$ has median 0.028 against 0.086; all six estimates share one sign in 80% of accepted against 35% of rejected triplets; the coefficient of variation is 0.30 against 1.09; and the magnitude of the mean of the six is the same in both groups (0.056 against 0.060), so the contrast is one of coherence, not of scale. Per treatment gene, the spread of the estimate across its accepted triplets is smaller than across its rejected triplets for 94% of the 78 genes with both (panel B). The contrast is not an artefact of the denominator guard of Appendix A.4: stratified by the quartile of the weakest relative denominator in the triplet, the within-triplet SD is 0.033 against 0.118 in the weakest quartile and 0.010 against 0.032 in the strongest. Two cautions. Agreement of the six estimates is algebraically implied by tetrad-compatibility of the same sample covariance, so it is a coherence and stability check of the accepted set rather than independent evidence of validity; and a triplet on the rank-one blind spot of Appendix B.5 would pass it. What it does rule out is that the accepted triplets owe their tetrad-compatibility to heteroskedastic or otherwise unstable covariance estimates: their six estimates agree where those of rejected triplets do not.

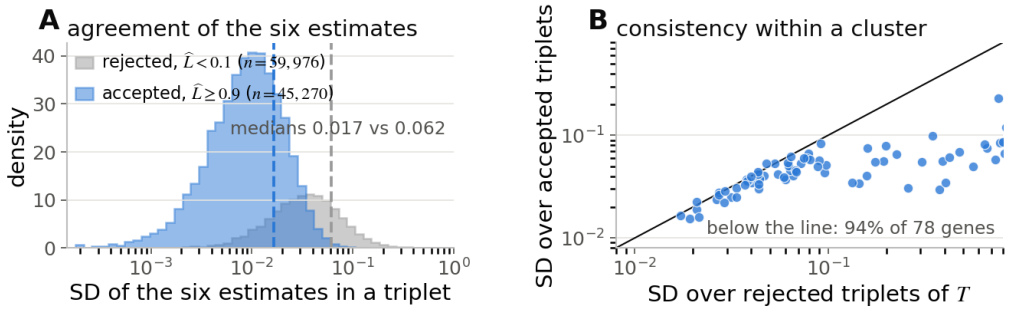


Figure 9: **Proxy-role symmetry in TCGA–BRCA. A:** within-triplet standard deviation of the six ordered-pair PCL estimates, accepted ($\widehat{L} \geq 0.9$, $n = 45{,}270$) against rejected ($\widehat{L} < 0.1$, $n = 59{,}976$) triplets; medians 0.017 and 0.062. **B:** for each treatment gene T, the standard deviation of its estimate across accepted triplets against that across rejected triplets; 94% of 78 genes lie below the identity.