Is correlation meaningfully correlated with causation? Crud, 1/f spectra, and a calibration standard for small-effect observational claims

Significance Statement

In large datasets from biomedicine, neuroscience, psychology, and genomics, nearly every pair of measured features is significantly correlated—the so-called "crud factor." These correlations are real, not artifacts of multiple testing: they reflect shared latent structure distributed broadly across many dimensions. We show that standard statistical adjustments cannot remove this background because shared variance in real-world data follows approximately power-law spectra, so no moderate-rank correction eliminates it. Many small associations reported as statistically significant are indistinguishable from this background, even after controlling for confounders. We prove that when an effect is comparable in size to the domain's background correlation scale, no purely statistical method can reliably separate signal from noise, regardless of sample size. We propose a practical diagnostic—a crud-aware z-test—that benchmarks any reported correlation against the domain background. Only associations that clearly exceed this background should be treated as evidence for a direct causal relationship.

Abstract

Causal claims in observational research increasingly rest on small adjusted correlations in large datasets. We quantify the background distribution of pairwise correlations across nine datasets spanning health, neuroscience, psychology, genomics, politics, and vision. After removing shared variation via principal component regression, the typical scale of residual correlations among random variable pairs—the "crud scale" σ_K—shrinks only slowly with the number of removed components. We trace this to approximately $1/f$ eigenvalue spectra, which distribute shared variance across many directions so that no moderate-dimensional adjustment can eliminate background dependence. A decision-theoretic result formalizes the consequence: when the causal signal is comparable to σ_K, no association-only rule can reliably distinguish direct causal relations from background dependence, regardless of sample size. We propose a crud-aware z-test that benchmarks observed correlations against the domain background. Associations that are highly significant by classical standards can become entirely unremarkable once background dependence is accounted for.

1 Introduction

Across biomedicine, social science, and behavioral research, causal claims increasingly rely on observational associations in large datasets. Randomized experiments remain the benchmark for counterfactual effects, but intervention is often infeasible, so small adjusted associations are routinely interpreted as evidence about causal mechanism. This paper asks how informative those

associations remain in the presence of the background dependence structure typical of modern datasets.

We organize interpretation around three uses of correlation: large effects that clear the crud scale easily, prediction without causal claims, and small-effect causal claims that depend on adjusted associations standing out from background dependence. Our results target *only* this third use. We do not argue that correlation is useless—large effects easily clear any background, and prediction tasks do not require causal interpretation. The concern is specifically with small adjusted associations that are interpreted as evidence for direct causal relationships.

Related work emphasizes that widespread weak dependence is common: Meehl's "crud factor" [Meehl, 1990] documented that in behavioral data, essentially everything correlates with everything. Sensitivity analyses [e.g., Cinelli and Hazlett, 2020, Oster, 2019] provide tools to benchmark how strong unobserved confounding would need to be to explain away an observed association. Their question is hypothetical: "how strong would confounding need to be to nullify this result?" Ours is empirical: "what does background dependence actually look like in this domain?" The two approaches are complementary, not competing. Separately, approximate $1/f$ spectral structure has been documented across many complex systems [Press, 1978, Keshner, 1982, Bak et al., 1987]. Our contribution is to connect these threads: we quantify the background correlation distribution across domains, link its slow shrinkage under generic adjustment to eigenvalue spectra, and propose a crud-aware calibration for interpreting small adjusted associations.

A concrete example previews the issue. A recent study reported a nominally significant association between folate intake and mortality in adults with type 2 diabetes ($r \approx 0.02$, $p < 0.05$, $n \approx 8{,}000$) [Liu et al., 2022]. But when we compare this association to the background distribution of correlations among NHANES variables—the "crud"—it falls squarely in the middle: its crud-aware p-value is 0.81, meaning the association is indistinguishable from a randomly chosen variable pair in the same dataset (Section 2.8). The classical test detects that the correlation is nonzero; it does not detect that the correlation is distinctive.

By "meaningfully correlated with causation," we mean the following narrow question. After a deliberately generic adjustment that removes broad shared variation (here operationalized via regression on the top K principal component scores), do the remaining pairwise associations separate from the domain's typical residual dependence? Specifically, are they large enough to support stable causal interpretation without additional design leverage (randomization, instruments, discontinuities, negative controls)? We operationalize the background dependence scale as the residual correlation standard deviation across uniformly sampled feature pairs and use it as the benchmark for whether an adjusted association is surprising. A formal bridge from the eigenvalue spectrum to this crud scale, and a decision-theoretic limit for association-only causal decisions, are given in Appendix A (Theorems 2–3). Across the nine domains we study, most small adjusted associations do not separate from the domain's background dependence.

Our approach treats the distribution of correlations among uniformly sampled feature pairs as a diagnostic of the domain's generic dependence structure. Under the common sparsity assumption that direct causal relations are rare relative to the number of measurable variables [cf. Spirtes et al., 2000, Ch. 2], most random pairs should not be directly connected, so the background correlation distribution is the relevant null for whether an observed association is distinctive. Concretely, we (i) quantify the empirical background correlation distribution across domains, (ii) measure how it changes under rank-K PC regression, (iii) document approximately power-law eigenvalue spectra, and (iv) propose evaluating associations against a crud-aware empirical null. The central contribution is the logic: if crud is ubiquitous and spectra are broad, then small adjusted associations are often not distinctive relative to the domain background; the bridge lemma formalizes this link.

2 Results

2.1 Datasets

We analyze nine large datasets spanning neuroscience, genomics, psychology, political science, vision, and population health (Figure 1): NHANES [National Center for Health Statistics, 2023], precinct-level 2016 voting returns merged with U.S. Census demographics [Hill et al., 2019], GTEx v8 RNA-seq [The GTEx Consortium, 2020], Allen Institute mouse brain RNA-seq [Tasic et al., 2018], HEXACO [Ashton and Lee, 2007], CIFAR-10 [Krizhevsky, 2009], Kay/Vim-1 fMRI [Kay et al., 2008], Haxby fMRI [Haxby et al., 2001], and Stringer mouse V1 [Stringer et al., 2019]. We focus on settings with large samples and many features because the dependence structure we study becomes easiest to characterize when sampling noise is small. These datasets allow us to measure the distribution of pairwise correlations across uniformly sampled feature pairs, how classical correlation significance behaves at typical sample sizes, and the eigenvalue spectrum that governs how much broad shared variation can be removed by low-rank adjustment.

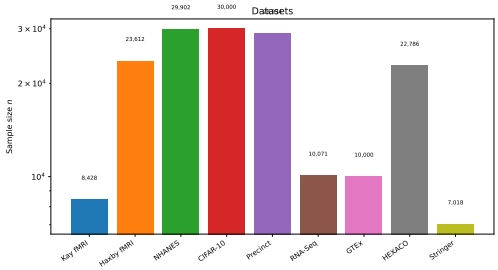

Figure 1: A set of large datasets used to study the correlational structure in typical modern science domains.

2.2 Empirical finding: broad correlations persist after adjustment

We first ask how correlated features typically are. If direct causal links are sparse, then uniformly sampled feature pairs are unlikely to be directly connected, so we would expect many correlations near zero after routine preprocessing. Instead, we observe a broad spread of correlations for random pairs (Figure 2, left; Table 1). Given the large sample sizes, classical iid-based correlation tests label a large fraction of these correlations as nominally significant, even though most random pairs are implausible candidates for direct causal relations. After PC regression (Methods; main figures use $K = 10$ with sensitivity across K in Table 1), the residual correlation distribution remains wide (Figure 2, right; Table 1). The core empirical fact is that the background correlation distribution is wide even after generic adjustment, so "significant" does not mean "distinctive." This is not a multiple-testing or "p is not effect size" point: even ignoring p-values entirely, the typical residual correlation scale after generic adjustment is substantial, so small adjusted associations are not distinctive evidence in the first place. Supplementary analyses compare PC regression to alternative adjustment strategies (e.g., covariate regression on randomly selected variables and sparse regression); results are reported in Appendix C.4. A counterexample would be a domain where a realistic adjustment used by practitioners collapses the residual correlation distribution, implying a rapidly decaying σ_K. We did not observe this in the domains studied here.

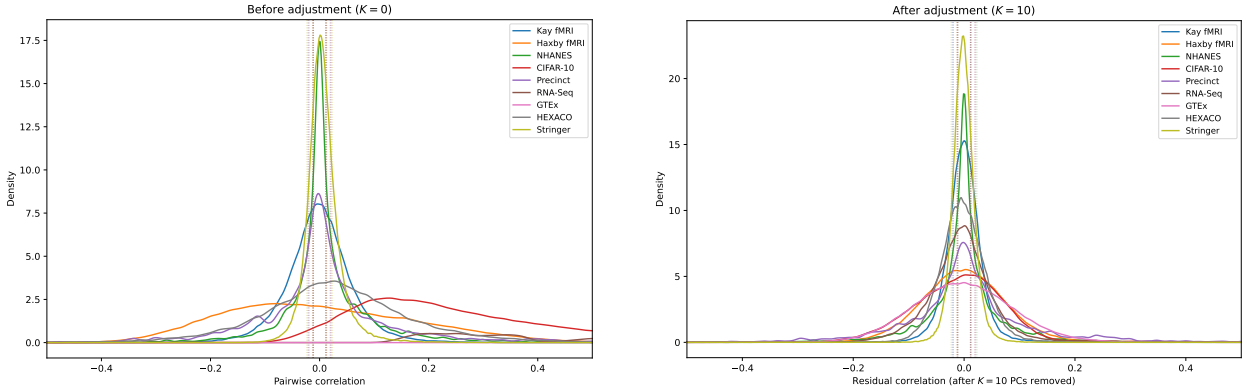

Figure 2: **Broad distributions of correlations persist after generic low-rank adjustment.** Left: distribution of pairwise correlations across uniformly sampled feature pairs before adjustment ($K = 0$). Right: the same distribution after regressing each feature on the top $K = 10$ principal component scores. Dotted vertical lines mark the classical two-sided significance threshold ($p < 0.05$) for each dataset's full sample size; nearly the entire crud distribution exceeds this threshold, illustrating that "significant" does not mean "distinctive."

2.3 Mechanism: power-law spectra

The persistence of substantial residual correlations after PC regression raises a basic question: why does removing broad shared variation not shrink the correlation distribution much faster? The answer lies in how shared variance is distributed across principal components. Each PC captures a fraction of the total variance; listing these fractions from largest to smallest gives the "eigenvalue spectrum." If the first few PCs captured nearly all the variance, removing them would eliminate most shared dependence. Instead, across datasets, explained variance decays approximately as a power law (Figure 3): on log–log axes the spectrum is roughly a straight line, meaning each successive PC captures a fixed fraction less variance than the previous one, so shared variation is distributed across many components rather than concentrated in the first few. This contrasts with the common implicit "spike" assumption behind low-rank adjustment: if confounding variance were concentrated in a few dominant components, removing them would collapse the background correlation distribution. Instead, removing the top K PCs eliminates only a modest portion of the dependence structure unless K is large, which is why residual correlations remain broadly distributed even at $K = 10$. Broad spectra make slow decay of the crud scale in K an expected outcome, not a surprise. Moreover, choosing a large K is not a remedy: with finite samples, principal components beyond the leading ones are estimated with increasing noise, so removing too many replaces bias with variance and can inflate residual correlations (as visible for Precinct at $K = 50$ in Table 1). Robustness checks for alternative explanations of the approximately power-law behavior are reported in Appendix C. In short, the background dependence in these domains cannot be removed by any low-rank adjustment of practical dimension.

2.4 Bridge lemma: the crud scale and its decay

The previous sections established that background correlations are wide and that eigenvalue spectra decay slowly. We now formalize the link: how wide should the residual correlation distribution be, given the eigenvalue spectrum? The answer is a simple formula—the "crud scale" σ_K—that predicts the spread of residual correlations from the spectrum alone.

Table 1: Off-diagonal correlation spread (SD) after removing the top K principal components. Entries report the standard deviation of off-diagonal correlations across uniformly sampled feature pairs. The spread generally decreases with K but remains nontrivial in all datasets. Non-monotonic increases at K=1 (CIFAR-10, GTEx) or K=50 (NHANES, Precinct) reflect redistribution of variance when a dominant component is removed or finite-p effects when K approaches the number of features.

Dataset	$K = 0$	$K = 1$	$K = 10$	$K = 50$
Kay fMRI	0.065	0.055	0.035	0.029
Haxby fMRI	0.175	0.150	0.077	0.038
NHANES	0.115	0.102	0.093	0.117
CIFAR-10	0.191	0.212	0.105	0.064
Precinct	0.166	0.159	0.136	0.293†
RNA-Seq	0.170	0.196	0.071	0.043
GTEx	0.038	0.255	0.090	0.043
HEXACO	0.144	0.125	0.046	0.040
Stringer	0.035	0.031	0.021	0.021

†Precinct has only p=83 features; removing 50 of 83 PCs leaves a near-degenerate residual subspace, inflating the residual correlation SD.

The key quantity is the width of the background correlation distribution after regressing out K principal components. In matrix notation, regressing each variable on the top K PC scores is equivalent to projecting out the top-K eigenspace of the covariance matrix Σ; the residual covariance is $\Sigma^{(K)} = (I - P_K)\Sigma(I - P_K)$, where P_K is the projector onto the leading K eigenvectors (details in Appendix A). The residual correlation between features i and j is the corresponding entry of $\Sigma^{(K)}$, normalized to unit diagonal: $\rho_{ij}^{(K)} = \Sigma_{ij}^{(K)} / \sqrt{\Sigma_{ii}^{(K)} \Sigma_{jj}^{(K)}}$.

Appendix A (Theorem 2) shows that for a uniformly random pair (i, j), the standard deviation of $\rho_{ij}^{(K)}$—i.e., how wide the residual correlation distribution is—is well approximated by a ratio of eigenvalue tail sums:

$$\sigma_K := \frac{\sqrt{\sum_{k>K} \lambda_k^2}}{\sum_{k>K} \lambda_k}.$$

Intuitively, the numerator measures how unevenly the remaining variance is distributed across components (the more concentrated it is, the wider the residual correlations), while the denominator measures the total remaining variance (which sets the scale). The result requires that individual PCs spread their influence broadly across features rather than loading on just a few—a condition called "eigenvector delocalization" that we verify empirically using standard diagnostics (Appendix B, Figures S1–S2). The spectral prediction tracks the empirical residual SD across values of K in all nine datasets.

In block-structured settings, σ_K should be read as a global crud scale; local or block-specific crud scales may be larger.

If the eigenvalues follow a power law $\lambda_k \approx ck^{-\alpha}$, then σ_K decays slowly as K grows (Appendix A, Corollary 1). In the especially relevant $\alpha \approx 1$ case,

$$\sigma_K \approx \frac{1}{\sqrt{K} \log(p/K)}.$$

Thus, even removing $K = 10$ components can leave a substantial residual crud scale when p is large. This formalizes the empirical observation that low-rank adjustment often reduces the spread

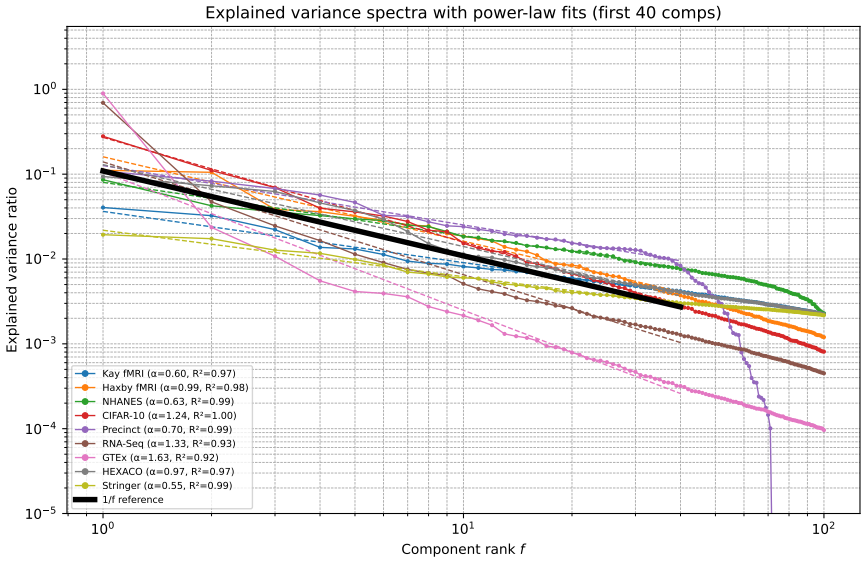

Figure 3: **Explained-variance spectra are approximately power-law, so correcting for shared variance is hard.** Each point shows how much variance a given PC explains; the axes are logarithmic (powers of 10). A straight line on these axes means a power law: variance decays as $k^{-\alpha}$, where the slope α controls how fast. When α is near 1 (as in most datasets here), variance is spread broadly, so removing the first few PCs leaves most of the shared dependence intact.

of correlations only modestly.

2.5 Decision-theoretic limit

The crud scale σ_K sets the width of the background correlation distribution. A natural follow-up question is: can we still reliably pick out truly causal pairs from this background? Standard significance testing asks whether an association differs from zero. The relevant question here is different: given that background correlations are not zero but instead drawn from a distribution of width σ_K, can any statistical rule reliably classify a pair as "causally linked" versus "merely correlated by background dependence"?

Appendix A (Theorem 3) formalizes this as a signal-detection problem. Consider a random pair (i, j): most pairs have no direct causal edge, and a small fraction have a direct edge that induces a mean shift μ in the adjusted association statistic T_{ij} (e.g., the adjusted sample correlation). Under the simplest form,

$$H_0 \text{ (no direct edge)}: \quad T_{ij} \sim \mathcal{N}(0, \sigma_K^2),$$
$$H_1 \text{ (direct edge)}: \quad T_{ij} \sim \mathcal{N}(\mu, \sigma_K^2).$$

Then the minimum achievable misclassification error for any decision rule that uses only T_{ij} is

$$\text{error}^* = \Phi\left(-\frac{|\mu|}{2\sigma_K}\right),$$

where Φ is the standard normal CDF. In particular, if $|\mu|$ is at most a constant multiple of σ_K, the error is bounded below by a constant bounded away from zero, regardless of sample size. This limitation persists as $n \to \infty$ because σ_K reflects population background dependence, not estimation error. Put plainly: if the causal signal (in adjusted-correlation units) is not larger than the crud

scale, then no correlation-only rule can reliably separate causal from noncausal pairs. If additional structure such as variance changes is also present under H_1, identification may become easier or harder depending on the direction of the change; the equal-variance mean-shift case provides the cleanest impossibility benchmark.

The base rate of direct causal edges makes this worse. In a gene expression study measuring $p = 20{,}000$ genes, there are roughly 200 million possible gene pairs, but each gene directly regulates only a handful of others, so the number of direct regulatory edges is on the order of p, not p^2. More generally, each newly measured variable has a bounded number of direct causes and effects, so the prior probability π that a randomly chosen pair shares a direct causal link is $O(1/p)$—vanishingly small in high-dimensional datasets. Sparse causal structure ($\pi \ll 1$) compounds the spectral limitation by collapsing positive predictive value even at moderate signal-to-crud ratios: not only is each per-pair decision noisy, but even "significant" pairs are overwhelmingly likely to be noncausal.

2.6 Crud-aware calibration

Normative implication. The minimum reporting standard for small-effect observational claims should be to compare any adjusted association to the background distribution produced by the same preprocessing and adjustment pipeline. Standard iid-based correlation tests ask whether an association differs from zero, not whether it is unusual relative to the domain's background correlations. In large datasets this produces a predictable failure mode: many associations have tiny p-values even when their magnitudes are typical for the domain. The relevant quantity is therefore a crud-aware percentile (or p-value) computed against a domain-calibrated empirical null. For a target pair (i, j) with an adjusted association statistic (for example, a residual correlation after removing K PCs), we compare its absolute value to the empirical distribution of absolute adjusted associations over uniformly sampled feature pairs from the same dataset, computed under the identical preprocessing and adjustment pipeline. The resulting crud-aware p-value is the fraction of random pairs whose absolute adjusted association exceeds the target's; the crud percentile is its complement (estimated by Monte Carlo sampling). The comparison is magnitude-based (two-sided): we rank by absolute value. For a uniformly random target pair processed through the same pipeline, this empirical p-value is (approximately) uniform by construction (Appendix A). For non-randomly chosen targets, it should be interpreted as a calibration score against the domain background rather than as a classical Type-I-error guarantee. Practically, reporting an adjusted association should include its percentile within the crud distribution; associations that do not exceed a high percentile (for example, the 95th or 99th) should be described as typical background dependence, even if their iid p-values are tiny.

Distinction from multiple testing. Unlike multiple-testing corrections, which assume a sharp null of zero effect, the crud percentile calibrates against a background of real nonzero dependencies induced by latent structure; it is a domain-calibrated baseline for interpretability, not an error-control device.

Stratified calibration. When dependence is heterogeneous, sample random pairs within the relevant comparison class (same modality, ROI, gene family, measurement type) so the baseline matches the target claim, and report which stratum was used.

2.7 A parametric crud-aware test for small studies

The empirical calibration above requires a data matrix large enough to estimate the background distribution of pairwise associations. Many studies, however, measure only a handful of variables on a small sample—a psychologist recruiting $n = 20$ participants and computing a single correlation, for instance. Such a study can be understood as observing a small submatrix of the much larger data matrix one *could* have collected had the budget permitted measuring all variables in the domain. The crud factor operates at the level of the full domain: if the typical background correlation among personality measures is $\sigma_{\text{crud}} \approx 0.15$, that background does not disappear simply because the investigator chose to measure only two of them.

The observed sample correlation r differs from the true population correlation ρ because of sampling noise, and ρ itself differs from zero because of background dependence. These are two independent sources of variability, and both contribute to the spread of r. Under a Gaussian approximation (Theorem 1 in Appendix A), the total variance of r for a random pair from the domain is the sum of these two contributions: $\text{Var}(r) \approx \sigma^2_{\text{crud}} + 1/(n-1)$, where σ^2_{crud} is the variance of true correlations across random pairs (the crud) and $1/(n-1)$ is the sampling variance for each pair. This yields a direct test statistic: for an observed correlation r with sample size n and domain crud level σ_{crud},

$$z_{\text{crud}} = \frac{|r|}{\sqrt{\sigma^2_{\text{crud}} + \frac{1}{n-1}}},$$

with two-sided p-value $p_{\text{crud}} = 2\,\Phi(-z_{\text{crud}})$. The classical correlation test uses $z_{\text{classical}} = |r|\sqrt{n-1}$, which is the special case $\sigma_{\text{crud}} = 0$: it tests against the null of zero population correlation, ignoring that a nonzero correlation may be entirely typical of the domain background.

When n is large, the sampling variance $1/(n-1)$ shrinks but the crud variance does not. A study with $n = 200$ and $r = 0.15$ yields $p < 0.05$ classically, yet $p_{\text{crud}} \approx 0.37$ when $\sigma_{\text{crud}} = 0.15$: the observed association is entirely typical of the domain background. In this regime, the classical test is detecting crud, not signal.

Diminishing returns beyond $n \approx 1/\sigma^2_{\text{crud}}$. The crud-aware framework also yields a concrete insight about sample size. The z-test denominator $\sqrt{\sigma^2_{\text{crud}} + 1/(n-1)}$ reveals a natural crossover. When n is small (specifically $n \ll 1/\sigma^2_{\text{crud}}$), sampling noise is the dominant source of uncertainty, so larger samples genuinely increase power to detect effects. But once n is large enough that the sampling noise $1/(n-1)$ becomes small relative to the crud variance σ^2_{crud}, the denominator is approximately σ_{crud} regardless of n: the bottleneck is no longer imprecise estimation but the fact that background correlations are themselves nonzero. Additional observations measure the crud more precisely but do not help separate signal from background. Because $\sigma_{\text{crud}} \approx 0.02$–$0.14$ across the domains in Table 1 (at $K = 10$), the crossover occurs at roughly $n \approx 50$–$2{,}500$. Beyond this point, classical p-values continue to shrink—explaining why massive studies routinely report "significant" but tiny effects—while the crud-aware p-value plateaus. For the purpose of distinguishing small associations from background dependence, there are sharply diminishing returns to collecting more than a few thousand observations per study.

How to use the crud-aware z-test.

1. Obtain the sample correlation r and sample size n from the study.
2. Look up a domain-appropriate crud scale σ_{crud}: use the residual correlation SD from a large reference dataset in your field (Table 1 reports values for various domains at $K = 10$), or consult published values.
3. Compute $z_{\text{crud}} = |r|/\sqrt{\sigma^2_{\text{crud}} + 1/(n-1)}$ and compare to standard normal critical values ($z > 1.96$ for two-sided $p < 0.05$). Report the crud-aware p-value alongside the classical one.

We recommend reporting conclusions for a range of plausible σ_{crud} values as a sensitivity analysis. When a large reference dataset is available, the full empirical calibration is preferable.

2.8 Worked examples

A large effect that clears the crud scale. In the HEXACO personality inventory ($p = 242$ items, $n = 100{,}000$), items within the same personality facet are designed to measure the same underlying trait. The median within-facet correlation is $|r| = 0.32$ at $K = 0$, and 55% of within-facet pairs fall in the top 5% of the crud distribution (11× enrichment over chance). Applying the crud-aware z-test with $\sigma_{\text{crud}} = 0.144$ (HEXACO at $K = 0$) gives $z_{\text{crud}} = 2.2$ for the median pair, $p_{\text{crud}} = 0.03$. After removing 10 PCs, 29% of within-facet pairs remain in the top 5% (6× enrichment). Known-strong associations survive the calibration.

A small effect that does not. In Liu et al. [2022], multivariable models report a nominally significant association between folate and mortality in adults with T2D (folate quartile 1 vs 2 HR 1.17, 95% CI 1.01–1.37) in an NHANES-derived cohort of adults with type 2 diabetes ($n \approx 8{,}000$). Converting from the reported confidence interval via $z = \log(\text{HR})/\text{SE}(\log \text{HR})$ and $r \approx z/\sqrt{n}$ gives $r \approx 0.02$, giving $p_{\text{classical}} \approx 0.04$. Applying the crud-aware z-test with $\sigma_{\text{crud}} = 0.093$ (NHANES at $K = 10$) gives $z_{\text{crud}} = 0.24$, $p_{\text{crud}} = 0.81$. Even this barely significant classical result becomes a crud-aware p-value of 0.81: the effect sits squarely in the middle of the domain's background correlation distribution, indistinguishable from a randomly chosen variable pair. For comparison, known strong biomarker pairs in the same dataset (hemoglobin vs hematocrit, $r = 0.91$; ALT vs AST, $r = 0.83$) easily clear the crud scale.

Additional worked examples are in SI Appendix C.5: NHANES biomarkers (positive—known mechanistic pairs clear the crud scale easily) and GTEx RNA-seq (negative—most reported cis-eQTL effects do not).

3 Discussion

As noted in the introduction, three common uses of correlation are: large effects that stand out far beyond the background dependence scale, prediction without causal claims, and small-effect causal claims that rely on adjusted associations standing out from background dependence. Our results target this third use: when reported adjusted associations are comparable to the crud scale, causal interpretation is fragile without additional design leverage, even if classical p-values are tiny.

The crud-aware calibration we propose provides a minimal standard for deciding when a small association is worth interpreting at all.

There are effects so large that pervasive background dependence is irrelevant. Smoking and lung cancer show extremely large associations; pooled multiple-adjusted relative risks are about 7 in large cohort meta-analyses, which easily clear any reasonable background dependence scale [O'Keeffe et al., 2018]. A second concrete example from a different domain is the HEXACO literature: Dark Triad traits are strongly negatively correlated with Honesty-Humility (rs $= -0.72, -0.57, -0.53$), which is far into the tail of any reasonable background correlation distribution [Lee and Ashton, 2005]. By contrast, across the modern domains we study, typical correlations are much weaker than these textbook examples, and many reported small effects do not clear the crud scale after generic adjustment. As a positive control, within-facet HEXACO item pairs clear the crud scale easily (best pair >99.99th percentile), showing the calibration identifies genuinely strong associations when they exist.

Scope and leverage. In some cases, prior knowledge or design-based leverage effectively isolates variables from background dependence. Randomized treatments, discontinuities, policy changes, and valid instruments provide leverage that is orthogonal to generic correlation structure. Similarly, informed selection of which pairs to examine—guided by prior biological knowledge, mechanistic models, or pathway databases—constitutes a structural prior that effectively narrows the background distribution; our impossibility result applies to the association statistic itself, not to the prior that selected the pair. Our results do not argue against such approaches; they target association-only workflows that interpret small adjusted pairwise correlations as evidence for direct causal structure in the absence of additional leverage. In our nine datasets, most adjusted correlations that would be called "significant" under iid testing fall near the center of the crud distribution rather than in its tails (Figure 2; Table 1), reinforcing the practical need for crud-aware calibration.

These observations also constrain automated causal reasoning: no amount of computational power can extract a signal that is not present in the data, so the crud scale bounds what any association-based system can learn, regardless of its sophistication (systems that combine data with mechanistic priors or interventions are not subject to this bound).

One striking upshot is the homogeneity of these properties across very different domains. The prevalence of broad spectra and slow shrinkage of the crud scale suggests that many modern datasets share a common statistical obstacle, and that calibration against background dependence is a cross-domain issue rather than a niche pathology.

3.1 Limitations

All results in this paper concern linear dependence: correlations, linear PC adjustment, and eigenvalue spectra of the covariance matrix. Many real-world domains exhibit substantial nonlinear dependence—gene regulatory networks involve thresholding and saturation, neural population codes are nonlinear functions of stimuli, and epidemiological dose-response curves are rarely linear. In such settings, the linear crud scale σ_K is a lower bound on the true background dependence: nonlinear shared variation that is invisible to PCA can still generate spurious pairwise associations in nonlinear statistics (e.g., mutual information, kernel-based dependence measures). This means our calibration is conservative with respect to nonlinear confounding—the actual background may be worse than what the linear analysis reveals.

Our empirical results are based on nine datasets chosen for size and availability, and the quantitative crud scale will vary across domains and preprocessing choices. We use PC regression as a

deliberately generic adjustment, not as a claim of causal identification. In high-dimensional settings, generic adjustment itself can be a source of uncertainty: when p is not small relative to n, estimating dominant directions and forming residualized associations can introduce nontrivial estimation error and instability, even when raw correlations are estimated precisely. This is most acute in datasets such as GTEx RNA-seq where $p \gg n$; in these settings, estimation of the eigenvalue tail and thus σ_K is noisier, and sensitivity checks are reported in Appendix C.4. These caveats do not change the main point, but they limit the precision of any crude calibration that ignores adjustment uncertainty.

Appendix

A Bridge lemmas: crud scale and limits of association-only causal inference

The main text documents an empirical pattern—broad background correlations that persist after low-rank adjustment—and proposes calibrating associations against this background. Here we formalize why the pattern arises and what it implies. We prove two results. First, the typical residual correlation across random feature pairs after removing K principal components is controlled by a simple ratio of eigenvalue tail sums (Theorem 2). When eigenvalues decay as a power law, this ratio shrinks only as $\sim 1/\sqrt{K}$, explaining the slow empirical decay in Table 1. Second, we show that any decision rule based solely on a one-dimensional adjusted association statistic incurs irreducible classification error when the causal signal is comparable to the crud scale (Theorem 3). Both proofs rely on standard random-matrix tools—Haar-like eigenvector moment bounds and Gaussian signal detection—and yield closed-form expressions that can be checked against the empirical values.

A.1 Setup and notation

Let $X \in \mathbb{R}^p$ be a mean-zero random vector with population covariance $\Sigma = \mathbb{E}[XX^\top] \in \mathbb{R}^{p\times p}$. Let $\Sigma = V\Lambda V^\top$ be an eigendecomposition with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. For an integer $K \in \{0, 1, \ldots, p-1\}$, let $P_K := \sum_{k=1}^{K} v_k v_k^\top$ be the projector onto the top-K eigenspace (with $P_0 = 0$). Define the population residual covariance after removing the top K principal components as

$$\Sigma^{(K)} := (I - P_K)\Sigma(I - P_K) = \sum_{k>K} \lambda_k v_k v_k^\top. \tag{1}$$

Define the population residual correlation for $i \neq j$ by

$$\rho_{ij}^{(K)} := \frac{\Sigma_{ij}^{(K)}}{\sqrt{\Sigma_{ii}^{(K)}\Sigma_{jj}^{(K)}}}. \tag{2}$$

A.2 Crud-aware p-values via random-pair calibration

A natural question is whether an observed adjusted association is unusual for the domain or merely typical background dependence. This subsection defines a "crud-aware" p-value that calibrates an observed adjusted association against the empirical background of the domain. The key idea is to treat the realized dataset as fixed and define a reference distribution by drawing feature pairs uniformly at random, while running the same preprocessing and adjustment pipeline as for

the target pair. The resulting calibration provides an assumption-light baseline for whether an observed association is distinctive.

Adjusted association statistic. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the observed data matrix (columns are features). Fix an adjustment procedure A_K (for example: regress each feature on the top K principal component scores and standardize the residuals). For any feature pair (a, b) with $1 \le a < b \le p$, define a one-dimensional adjusted association statistic

$$T_{ab} = T(\mathbf{X}; a, b, K). \tag{3}$$

A canonical example is the sample correlation of adjusted (residualized) features, $T_{ab} = \widehat{\rho}_{ab}^{(K)}$. The only requirement is that the same pipeline (preprocessing, adjustment, and statistic) is used for the target pair and for reference pairs.

Crud-null reference distribution. Rather than testing whether T_{ij} is zero, we compare its magnitude to what is typical for a uniformly random feature pair under the same pipeline. Let

$$(A, B) \sim \mathrm{Unif}\Big(\{(a, b) : 1 \le a < b \le p\}\Big), \tag{4}$$

independent of all other randomness. For a threshold $t \ge 0$, define the (two-sided) crud-null tail probability

$$p_{\mathrm{crud}}(t) \;:=\; \mathbb{P}\big(|T_{AB}| \ge t \,\big|\, \mathbf{X}\big). \tag{5}$$

This is a conditional probability given the realized dataset $\mathbf{X}$, where the only randomness is the selection of a uniformly random feature pair.

Crud-aware p-value for a target pair. For a specific target pair (i, j), set $t_{\mathrm{obs}} = |T_{ij}|$ and define

$$p_{\mathrm{crud}}(i, j) \;:=\; \mathbb{P}\big(|T_{AB}| \ge |T_{ij}| \,\big|\, \mathbf{X}\big). \tag{6}$$

Equivalently, $p_{\mathrm{crud}}(i, j)$ is the fraction of feature pairs whose absolute adjusted association equals or exceeds that of the target pair—a two-sided, magnitude-based rank. The *crud percentile* is the complement $1 - p_{\mathrm{crud}}(i, j)$: values near 1 indicate that the target pair's association is unusually large relative to the domain background.

Monte Carlo estimation. In practice, $p_{\mathrm{crud}}(i, j)$ is approximated by sampling M reference pairs uniformly from all $\binom{p}{2}$ pairs, computing $T_{A_m B_m}$ for each, and taking the fraction (plus one, for the standard add-one correction) whose absolute value exceeds $|T_{ij}|$. For a uniformly random target pair, p_{crud} is (super-)uniform by a standard exchangeability argument. For non-randomly chosen targets, it should be interpreted as a domain-calibrated extremeness score rather than a Type-I-error guarantee.

Practical refinements. When features are heterogeneous in type or scale, reference pairs should be stratified (e.g., within the same modality, ROI, or gene family) so the baseline matches the target claim. When crud-aware p-values are computed for many pre-specified pairs, standard multiplicity corrections (e.g., Benjamini–Hochberg) can be applied. The definition of p_{crud} depends only on the choice of adjustment pipeline and statistic; alternative adjustment strategies are compared in Appendix C.4.

A parametric crud-aware test for small studies. The empirical calibration above requires a data matrix large enough to estimate the background distribution of pairwise associations. Many studies, however, measure only a handful of variables on a small sample—a psychologist recruiting $n = 20$ participants and computing a single correlation, for instance. Such a study can be understood as observing a small submatrix of the much larger data matrix one *could* have collected had the budget permitted measuring all variables in the domain. The crud factor operates at the level of the full domain: if the typical background correlation among personality measures is $\sigma_{\text{crud}} \approx 0.15$, that background does not disappear simply because the investigator chose to measure only two of them. This motivates a parametric analogue of the crud-aware p-value that requires only summary statistics.

Setup. Suppose an investigator observes the sample correlation $\hat{\rho}$ between two variables measured on n independent observations. These two variables are drawn from a domain of p variables with population covariance Σ. Define the population correlation for the pair (i, j) as $\rho_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, and let σ^2_{crud} denote the variance of ρ_{ij} when (i, j) is drawn uniformly at random from all $\binom{p}{2}$ pairs. The quantity σ_{crud} characterizes the spread of background correlations in the domain and can be estimated from large published datasets or reference analyses. (When K principal components have been removed, one uses the residual correlations $\rho^{(K)}_{ij}$ and the corresponding $\sigma^{(K)}_{\text{crud}}$.)

Theorem 1 (Parametric crud-aware test)**.** *Suppose the pair (I, J) is drawn uniformly at random from $\{(a, b) : 1 \le a < b \le p\}$, and conditionally on (I, J), the sample correlation $\hat{\rho}$ satisfies*

$$\hat{\rho} = \rho_{IJ} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2), \qquad \sigma_n^2 = \frac{1 - \rho_{IJ}^2}{n - 1} \approx \frac{1}{n - 1}, \tag{7}$$

where ε is independent of (I, J) and the approximation $\sigma_n^2 \approx 1/(n - 1)$ holds when $|\rho_{IJ}|$ is small. Further suppose that the population correlation ρ_{IJ}, viewed as a random variable over the uniform draw of pairs, satisfies

$$\rho_{IJ} \sim \mathcal{N}(0, \sigma^2_{\text{crud}}). \tag{8}$$

Then the marginal distribution of the sample correlation under the crud null is

$$\hat{\rho} \sim \mathcal{N}\left(0,\ \sigma^2_{\text{crud}} + \frac{1}{n - 1}\right). \tag{9}$$

Proof. Write $\hat{\rho} = \rho_{IJ} + \varepsilon$. By assumption, $\rho_{IJ} \sim \mathcal{N}(0, \sigma^2_{\text{crud}})$ and $\varepsilon \sim \mathcal{N}(0, 1/(n - 1))$, and these are independent (the sampling noise ε is conditionally independent of the pair identity given ρ_{IJ}, and under the Gaussian approximation in (7) with $\sigma_n^2 \approx 1/(n - 1)$, the conditional variance does not depend on ρ_{IJ}). The sum of two independent Gaussians is Gaussian with variance equal to the sum of variances:

$$\text{Var}(\hat{\rho}) = \text{Var}(\rho_{IJ}) + \text{Var}(\varepsilon) = \sigma^2_{\text{crud}} + \frac{1}{n - 1}. \tag{10}$$

Both components have zero mean, so $\mathbb{E}[\hat{\rho}] = 0$, yielding (9). □

The crud-aware z-test. Theorem 1 yields a direct test statistic. For an observed correlation r with sample size n and domain crud level σ_{crud}, define

$$z_{\text{crud}} = \frac{|r|}{\sqrt{\sigma^2_{\text{crud}} + \frac{1}{n-1}}}, \tag{11}$$

with two-sided p-value $p_{\text{crud}} = 2\,\Phi(-z_{\text{crud}})$. For comparison, the classical correlation test uses $z_{\text{classical}} = |r|\sqrt{n - 1}$, which is the special case $\sigma_{\text{crud}} = 0$: it tests against the null of zero population correlation, ignoring that a nonzero correlation may be entirely typical of the domain background.

When the correction matters. The ratio of the two test statistics is

$$\frac{z_{\text{crud}}}{z_{\text{classical}}} = \frac{1}{\sqrt{1 + (n-1)\,\sigma_{\text{crud}}^2}}. \tag{12}$$

When n is small relative to $1/\sigma_{\text{crud}}^2$, this ratio is close to 1 and the crud correction is modest—though it still raises the significance threshold, reflecting that a correlation of $r = 0.3$ in a field with $\sigma_{\text{crud}} = 0.15$ is less remarkable than the same r in a field with negligible background. When n is large, the sampling variance $1/(n-1)$ shrinks but the crud variance does not. A study with $n = 200$ and $r = 0.15$ yields $p < 0.05$ classically, yet $p_{\text{crud}} \approx 0.37$ when $\sigma_{\text{crud}} = 0.15$: the observed association is entirely typical of the domain background. In this regime, the classical test is detecting crud, not signal.

Assumptions and practical guidance. The derivation relies on two approximations. First, the Gaussian model (8) for the distribution of population correlations over random pairs. This is supported by the empirical observation (Section A.2 and the normality audit in the main text) that residual correlations are well-approximated by a Gaussian after removing top principal components. Second, the linearization $\sigma_n^2 \approx 1/(n-1)$, which is standard for $|\rho|$ not too close to ± 1 and n not too small.

In practice, σ_{crud} must be estimated from external data, introducing additional uncertainty. We recommend reporting the conclusion for a range of plausible σ_{crud} values as a sensitivity analysis. When a large reference dataset is available, the full empirical calibration of Section A.2 is preferable. The parametric version is intended for the common setting where only summary statistics are at hand—a simple corrective to the standard practice of testing against the independence null alone.

A.3 Eigenvector delocalization assumption

To connect the eigenvalue spectrum to typical pairwise correlations, we need that eigenvectors are not concentrated on small sets of coordinates.

Assumption 1 (Haar-like eigenvectors / delocalization)**.** *The eigenvector matrix V is distributed as a Haar-uniform random orthogonal matrix, independent of Λ, or (more generally) satisfies the moment bounds of Haar orthogonal matrices: for all $i \neq j$ and all k,*

$$\mathbb{E}[v_{ik}] = 0, \qquad \mathbb{E}[v_{ik}^2] = \frac{1}{p}, \qquad \mathbb{E}[v_{ik}^2 v_{jk}^2] = \frac{1}{p(p+2)}. \tag{13}$$

Additionally, cross-component correlations are negligible in the sense that the covariance of $v_{ik}v_{jk}$ and $v_{i\ell}v_{j\ell}$ for $k \neq \ell$ contributes only lower-order terms ($O(p^{-3})$) to the variance calculations below.

Assumption 1 is standard in random matrix theory as a proxy for "incoherent" or "delocalized" eigenvectors. Empirically, it is supported when no small subset of variables dominates principal components.

Remark (localized crud). If eigenvectors are partially localized or block-structured, the same calculations imply that crud may be localized rather than global: typical residual correlations can remain large within blocks even if global random-pair correlations are smaller.

A.4 Theorem A.1: crud scale after removing top-K principal components

Theorem 2 (Crud scale under top-K PC removal)**.** *Assume Assumption 1. Fix $K \in \{0, \dots, p-1\}$ and define tail sums*

$$S_1(K) := \sum_{k>K} \lambda_k, \qquad S_2(K) := \sum_{k>K} \lambda_k^2. \tag{14}$$

Then for a uniformly random pair (i, j) with $i \neq j$,

$$\mathrm{SD}\Big(\rho_{ij}^{(K)}\Big) = \frac{\sqrt{S_2(K)}}{S_1(K)} \cdot \big(1 + o_p(1)\big), \tag{15}$$

as $p \to \infty$ (with K allowed to grow with p, provided $K < p$ and the tail is nontrivial). Equivalently, the typical residual correlation magnitude across random feature pairs is controlled by

$$\sigma_K := \frac{\sqrt{\sum_{k>K} \lambda_k^2}}{\sum_{k>K} \lambda_k}. \tag{16}$$

Proof. Write the off-diagonal residual covariance entry as

$$\Sigma_{ij}^{(K)} = \sum_{k>K} \lambda_k v_{ik} v_{jk}, \qquad i \neq j. \tag{17}$$

By symmetry under Assumption 1, $\mathbb{E}[v_{ik}v_{jk}] = 0$, hence $\mathbb{E}[\Sigma_{ij}^{(K)}] = 0$.

For the variance, expand

$$\mathrm{Var}(\Sigma_{ij}^{(K)}) = \sum_{k>K} \lambda_k^2 \,\mathrm{Var}(v_{ik}v_{jk}) + 2 \sum_{K<k<\ell} \lambda_k \lambda_\ell \,\mathrm{Cov}(v_{ik}v_{jk}, v_{i\ell}v_{j\ell}). \tag{18}$$

Under the Haar moment bounds,

$$\mathrm{Var}(v_{ik}v_{jk}) = \mathbb{E}[v_{ik}^2 v_{jk}^2] - \mathbb{E}[v_{ik}v_{jk}]^2 = \frac{1}{p(p+2)}. \tag{19}$$

The cross-component covariance term is of smaller order (Assumption 1), so

$$\mathrm{Var}(\Sigma_{ij}^{(K)}) = \frac{1}{p(p+2)} \sum_{k>K} \lambda_k^2 \cdot \big(1 + o(1)\big), \tag{20}$$

and therefore

$$\mathrm{SD}(\Sigma_{ij}^{(K)}) = \frac{1}{p}\sqrt{S_2(K)} \cdot \big(1 + o(1)\big). \tag{21}$$

Next, consider the diagonal:

$$\Sigma_{ii}^{(K)} = \sum_{k>K} \lambda_k v_{ik}^2. \tag{22}$$

Taking expectations gives

$$\mathbb{E}[\Sigma_{ii}^{(K)}] = \sum_{k>K} \lambda_k \mathbb{E}[v_{ik}^2] = \frac{1}{p} S_1(K). \tag{23}$$

Under Assumption 1, $\Sigma_{ii}^{(K)}$ concentrates around its mean with relative fluctuations $o_p(1)$ (a standard consequence of bounded fourth moments and orthogonality). Hence

$$\sqrt{\Sigma_{ii}^{(K)}\Sigma_{jj}^{(K)}} = \frac{1}{p} S_1(K) \cdot \big(1 + o_p(1)\big). \tag{24}$$

Finally, for $i \neq j$,

$$\rho_{ij}^{(K)} = \frac{\Sigma_{ij}^{(K)}}{\sqrt{\Sigma_{ii}^{(K)}\Sigma_{jj}^{(K)}}} = \frac{\Sigma_{ij}^{(K)}}{\frac{1}{p}S_1(K)} \cdot \big(1 + o_p(1)\big). \tag{25}$$

Taking standard deviations and substituting the expression for $\mathrm{SD}(\Sigma_{ij}^{(K)})$ yields

$$\mathrm{SD}(\rho_{ij}^{(K)}) = \frac{\frac{1}{p}\sqrt{S_2(K)}}{\frac{1}{p}S_1(K)} \cdot \big(1 + o_p(1)\big) = \frac{\sqrt{S_2(K)}}{S_1(K)} \cdot \big(1 + o_p(1)\big). \tag{26}$$

□

Corollary 1 (Power-law eigenvalues imply slow decay in K)**.** *Assume $\lambda_k = ck^{-\alpha}$ for $k = 1, \ldots, p$ with $c > 0$ and $\alpha > 1/2$, and assume the conditions of Theorem 2. Then for $1 \ll K \ll p$, the crud scale $\sigma_K = \sqrt{S_2(K)}/S_1(K)$ satisfies:*

- *If $\alpha = 1$, then $\sigma_K \asymp \dfrac{1}{\sqrt{K}\,\log(p/K)}$.*
- *If $\alpha > 1$, then $\sigma_K \asymp \dfrac{1}{\sqrt{K}}$.*
- *If $1/2 < \alpha < 1$, then $\sigma_K \asymp \dfrac{1}{\sqrt{K}}\left(\dfrac{K}{p}\right)^{1-\alpha}$.*

Here $a_K \asymp b_K$ means a_K/b_K is bounded above and below by positive constants depending only on (c, α).

Proof. By Theorem 2, it suffices to compute asymptotics of $S_1(K)$ and $S_2(K)$. Approximate sums by integrals.

If $\alpha = 1$: $S_1(K) = c\sum_{k=K+1}^{p} k^{-1} \asymp c\log(p/K)$ and $S_2(K) = c^2\sum_{k=K+1}^{\infty} k^{-2} \asymp c^2/K$. Thus $\sigma_K \asymp (c/\sqrt{K})/(c\log(p/K))$.

If $\alpha > 1$: $S_1(K) \asymp c\int_K^{\infty} x^{-\alpha}dx \asymp cK^{1-\alpha}$ and $S_2(K) \asymp c^2\int_K^{\infty} x^{-2\alpha}dx \asymp c^2K^{1-2\alpha}$, so $\sigma_K \asymp K^{-1/2}$.

If $1/2 < \alpha < 1$: $S_1(K) \asymp c\int_K^{p} x^{-\alpha}dx \asymp c\,(p^{1-\alpha} - K^{1-\alpha}) \asymp c\,p^{1-\alpha}$ and $S_2(K) \asymp c^2K^{1-2\alpha}$, giving $\sigma_K \asymp (cK^{(1-2\alpha)/2})/(cp^{1-\alpha}) = \frac{1}{\sqrt{K}}\left(\frac{K}{p}\right)^{1-\alpha}$. □

A.5 Theorem A.2: limits of association-only causal decisions

We now formalize a limitation of procedures that treat small adjusted associations as evidence for direct causation.

Theorem 3 (Association-only decisions cannot be reliable below the crud scale)**.** *Fix K and consider a uniformly random feature pair (i, j) with $i \neq j$. Suppose the analyst compresses the data to a one-dimensional adjusted association statistic T_{ij} (for example $T_{ij} = \hat{\rho}_{ij}^{(K)}$) and uses a decision rule $\delta : \mathbb{R} \to \{0, 1\}$ where $\delta(T_{ij}) = 1$ means "direct causal edge present".*

Assume a stylized mixture model:

$$H_0 \text{ (no direct edge)}: \quad T_{ij} \sim N(0, \sigma_K^2), \tag{27}$$

$$H_1 \text{ (direct edge)}: \quad T_{ij} \sim N(\mu, \sigma_K^2), \tag{28}$$

where σ_K is the crud scale from Theorem 2 and μ is the mean shift induced by the direct causal effect after generic adjustment.

Then (for equal prior probabilities on H_0 and H_1) the minimum achievable misclassification error is

$$\inf_{\delta} \ \mathbb{P}\big(\delta(T_{ij}) \text{ is wrong}\big) = \Phi\left(-\frac{|\mu|}{2\sigma_K}\right). \tag{29}$$

In particular, if $|\mu| \le c\,\sigma_K$, then

$$\inf_{\delta} \ \mathbb{P}\big(\delta(T_{ij}) \text{ is wrong}\big) \ \ge \ \Phi(-c/2), \tag{30}$$

a constant bounded away from 0 independent of sample size.

Proof. Under the stated model, T_{ij} follows one of two normal distributions with equal variance: $N(0, \sigma_K^2)$ under H_0 and $N(\mu, \sigma_K^2)$ under H_1. By the Neyman–Pearson lemma, the likelihood ratio test is optimal. For equal priors and $\mu > 0$, the optimal threshold is $t^\star = \mu/2$ (symmetrically $-\mu/2$ for $\mu < 0$). Then

$$\mathbb{P}(\text{error} \mid H_0) = \mathbb{P}(T_{ij} > \mu/2 \mid H_0) = 1 - \Phi\left(\frac{\mu}{2\sigma_K}\right), \tag{31}$$

$$\mathbb{P}(\text{error} \mid H_1) = \mathbb{P}(T_{ij} \le \mu/2 \mid H_1) = \Phi\left(-\frac{\mu}{2\sigma_K}\right). \tag{32}$$

These are equal, so the Bayes error equals $\Phi(-\mu/(2\sigma_K))$. Replacing μ by $|\mu|$ handles both signs. The second inequality follows immediately if $|\mu| \le c\sigma_K$. □

Remark (the bound is a benchmark, not always a lower bound). The equal-variance Gaussian model provides the cleanest impossibility benchmark. If additional structure such as variance changes is also present under H_1, identification may become easier (if signal pairs have distinctive variance) or harder (if the variance change increases overlap). The mean-shift formulation isolates the most basic obstruction: when the causal signal in adjusted-correlation units is comparable to the background scale, no thresholding rule on T_{ij} alone can reliably separate signal from noise.

Remark (rare edges make the problem worse). If the prior probability of a direct causal edge is $\pi \ll 1$, then even when $|\mu|/\sigma_K$ is moderate, the positive predictive value of any thresholding rule can remain small. Theorem 3 isolates the more basic point: when $|\mu|$ is on the order of the crud scale, no association-only rule can achieve low error.

Remark (robustness of the limit). The Gaussian mixture model is a stylized proxy for association-only workflows that compress each pair to a one-dimensional statistic. The qualitative limit depends on signal-to-noise: if the causal signal is not large relative to the background scale, discrimination is poor. Unequal variances or heavier-tailed distributions only make separation harder, and any rule that discards multivariate structure is subject to the same basic limitation.

A.6 Robustness of the assumptions

The theorems above use idealized assumptions—Haar-like eigenvectors, diagonal concentration, and a Gaussian mixture—to obtain closed-form expressions. Real datasets exhibit moderate eigenvector

localization, finite-p effects, and heavy-tailed residuals. These deviations mainly affect constants and tail behavior, not the qualitative conclusions. When eigenvectors are partially localized, the spectrum-only formula in Theorem 2 typically becomes conservative (localization increases diagonal heterogeneity, reducing average off-diagonal energy). The Gaussian error curve in Theorem 3 is likewise a benchmark: heavy-tailed backgrounds only increase overlap for a fixed mean shift. Appendix B audits these assumptions empirically across all nine datasets and confirms that the spectral prediction tracks the observed residual SD and that qualitative conclusions are robust.

B Empirical audit of assumptions

Theorems 2 and 3 rely on idealized assumptions; the question is whether these assumptions hold well enough in practice to support the qualitative conclusions. We test this on nine real-world datasets spanning diverse domains: Kay fMRI (visual neuroscience, n=8428, p=1870), Haxby fMRI (cognitive neuroscience, n=23612, p=1452), Stringer neural recordings (mouse cortex, n=7018, p=6000), NHANES (population health, n=29902, p=165), CIFAR-10 images (n=30000, p=3072), precinct-level voting and census demographics (n=28934, p=83), RNA-Seq gene expression (Allen Institute mouse brain, n=10071, p=387), GTEx gene expression (human skeletal muscle, n=10000, p=803), and HEXACO personality inventory (n=22786, p=242). All datasets are z-scored by feature before analysis; up to n=10000 samples are used per dataset. PCA is performed via randomized SVD with up to 300 components. For each test, we report both a diagnostic summary and an interpretation of what deviations from the ideal mean for the qualitative message.

B.1 Test 1: Power-law eigenvalue decay

Theorem 2 and Corollary 1 assume that eigenvalues follow an approximate power law $\lambda_k \propto k^{-\alpha}$. We fit $\log(\lambda_k/\text{total var}) = a - \alpha \log k$ on the first 40 components and compare to an exponential model via AIC.

Table 2: Power-law fits on first 40 principal components.

Dataset	α	R^2	AIC (power law)	AIC (exponential)	Preferred
NHANES	0.63	0.992	−237	−121	power law
Kay fMRI	0.67	0.984	−204	−97	power law
Precinct	0.70	0.985	−205	−105	power law
Stringer	0.74	0.987	−206	−109	power law
HEXACO	0.97	0.974	−155	−70	power law
Haxby fMRI	1.03	0.982	−165	−98	power law
CIFAR-10	1.24	0.997	−218	−64	power law
GTEx	1.30	0.994	−190	−62	power law
RNA-Seq	1.33	0.932	−90	−24	power law

All nine datasets are better described by a power law than an exponential over the first 40 components ($\Delta\text{AIC} > 60$ in every case). Exponents range from $\alpha \approx 0.63$ (NHANES) to $\alpha \approx 1.33$ (RNA-Seq), placing all datasets in the regime $\alpha > 1/2$ required by Corollary 1. Most datasets cluster near $\alpha \approx 0.7$–1.0, consistent with $1/f$-like spectral decay.

B.2 Test 2: Eigenvector delocalization

Assumption 1 requires that eigenvectors are not concentrated on small subsets of coordinates. We measure this via coherence $\mu := \max_{i,k} |v_{ik}|$ (ideal: $1/\sqrt{p}$ under Haar), the inverse participation ratio $\mathrm{IPR}(k) := \sum_i v_{ik}^4$ (ideal: $1/p$), and leverage score uniformity.

Table 3: Eigenvector delocalization diagnostics (top 50 PCs).

Dataset	Coherence	Ideal $1/\sqrt{p}$	Ratio	Med. eff. support	p
Stringer	0.082	0.013	6.4×	1542	6000
GTEx	0.194	0.035	5.5×	245	803
CIFAR-10	0.069	0.018	3.8×	1216	3072
Kay fMRI	0.118	0.023	5.1×	556	1870
Haxby fMRI	0.131	0.026	5.0×	478	1452
RNA-Seq	0.572	0.051	11.2×	57	387
HEXACO	0.371	0.064	5.8×	73	242
NHANES	0.439	0.078	5.6×	29	165
Precinct	0.908	0.110	8.3×	12	83

The degree of delocalization varies systematically with p. High-dimensional datasets (Stringer, CIFAR-10, GTEx, Kay, Haxby) show moderate delocalization with coherence ratios of 3.8–6.4× ideal and effective supports spanning 25–50% of p. Low-dimensional datasets (Precinct, NHANES) exhibit substantial localization, with coherence ratios up to 8.3× and effective supports well below p. RNA-Seq shows the most localization (11.2×), consistent with the presence of highly co-regulated gene modules. As discussed in Section A.6, localization typically makes the spectrum-only formula *conservative* (overestimates typical correlations), so the qualitative conclusions of Theorem 2 remain valid even where eigenvectors are not perfectly delocalized.

B.3 Test 3: Predicted σ_K versus empirical residual correlation SD

The central prediction of Theorem 2 is that the standard deviation of off-diagonal residual correlations after removing K PCs is approximately $\sigma_K = \sqrt{S_2(K)}/S_1(K)$. We compare this spectral prediction to the empirically measured SD of residual correlations on a random subset of 500 features.

Table 4: Correlation between predicted σ_K and empirical residual correlation SD across $K \in \{0, 1, 2, 5, 10, 20, 50, 100\}$.

Dataset	Pearson r (predicted vs. empirical)
HEXACO	0.997
Precinct	0.997
Haxby fMRI	0.996
NHANES	0.987
GTEx	0.987
Stringer	0.952
Kay fMRI	0.942
CIFAR-10	0.880
RNA-Seq	0.639

In seven of nine datasets, the spectral prediction tracks the empirical SD with $r > 0.94$. CIFAR-10 ($r = 0.88$) and RNA-Seq ($r = 0.64$) show weaker agreement, the former due to strong eigenvector structure (spatial image statistics) and the latter due to gene-module localization. Even where the spectral formula is quantitatively imprecise, the monotonic relationship between K and the crud scale is preserved, supporting the qualitative message that broad spectra produce slowly decaying background correlations.

B.4 Test 4: Diagonal concentration of residual covariance

The proof of Theorem 2 requires that the diagonal entries $\Sigma_{ii}^{(K)}$ concentrate around their mean $S_1(K)/p$. We measure this via the coefficient of variation (CV) of the residual variance $\mathrm{Var}_n(X_i^{(K)})$ across features i.

Table 5: Coefficient of variation of residual variances across features.

Dataset	K=0	K=1	K=10	K=50
Stringer	0.000	0.036	0.091	0.136
CIFAR-10	0.000	0.217	0.222	0.216
Kay fMRI	0.000	0.052	0.105	0.167
Haxby fMRI	0.000	0.122	0.149	0.163
HEXACO	0.000	0.097	0.171	0.220
GTEx	0.000	0.303	0.516	0.689
NHANES	0.177	0.234	0.456	0.811
RNA-Seq	0.000	0.550	0.544	0.320
Precinct	0.000	0.177	0.589	1.127

The pattern is clear: diagonal concentration holds well for large p and degrades predictably as p shrinks and K grows. At K=0 (before PC removal), z-scoring enforces unit variance by construction (CV ≈ 0). As PCs are removed, diagonal heterogeneity increases. High-p datasets (Stringer, Kay, Haxby) maintain good concentration (CV < 0.2) even at K=50. Low-p datasets (Precinct, NHANES) show substantial heterogeneity at large K, consistent with the finite-p discussion in Section A.6. This diagonal heterogeneity makes the Haar-based formula conservative, so the spectrum-only prediction is an upper bound on typical off-diagonal correlations in these settings.

B.5 Test 5: Normality of residual correlations

Theorem 3 models the null distribution of residual associations as Gaussian. We assess this by computing the skewness and excess kurtosis of the off-diagonal residual correlations.

Residual correlations are generally right-skewed with positive excess kurtosis (heavier tails than Gaussian). GTEx and Haxby fMRI at K=0 are the closest to Gaussian (skew $-0.26/0.33$, kurtosis $-0.35/-0.34$). NHANES shows the largest departures. As noted in Section A.6, heavy tails increase the overlap between null and alternative distributions for a given mean shift, so the Gaussian error formula in Theorem 3 is optimistic about separability—the true error is at least as large.

B.6 Test 6: Cross-component covariance negligibility

Assumption 1 requires that the covariance of $v_{ik}v_{jk}$ and $v_{i\ell}v_{j\ell}$ for $k \neq \ell$ is negligible. We estimate this by computing the lag-1 autocorrelation of the product sequence $\{v_{ik}v_{jk}\}_{k=1}^{K}$ across 1000 random variable pairs (i, j).

Table 6: Skewness and excess kurtosis of off-diagonal residual correlations.

	K=0		K=10	
Dataset	Skew	Kurt	Skew	Kurt
HEXACO	0.07	1.94	0.26	17.36
GTEx	−0.26	−0.35	0.30	1.41
Haxby fMRI	0.33	−0.34	0.17	1.34
Kay fMRI	0.39	5.46	1.45	24.07
CIFAR-10	0.99	0.63	2.47	13.99
Precinct	1.25	8.44	0.86	8.87
RNA-Seq	−1.61	1.96	1.98	12.61
Stringer	1.98	14.49	1.23	11.00
NHANES	2.99	21.43	2.90	35.58

Table 7: Cross-component covariance: autocorrelation of eigenvector products.

Dataset	Mean cross-corr	|mean|	Frac. $\|r\| > 0.1$
RNA-Seq	−0.013	0.013	0.493
CIFAR-10	−0.018	0.018	0.469
Kay fMRI	−0.019	0.019	0.467
GTEx	−0.027	0.027	0.483
Stringer	−0.025	0.025	0.466
Haxby fMRI	−0.024	0.024	0.459
HEXACO	−0.028	0.028	0.497
NHANES	−0.035	0.035	0.552
Precinct	−0.061	0.061	0.568

The mean cross-component correlation is small in magnitude (< 0.07) for all datasets, with higher-p datasets showing smaller values as expected. While roughly half of individual pairs show $|r| > 0.1$, these are centered near zero and largely cancel in the variance calculation, consistent with the $O(p^{-3})$ bound in Assumption 1.

B.7 Summary of assumption audit

The assumptions of Theorems 2 and 3 hold to a reasonable approximation across all nine datasets, with the quality of the approximation improving systematically with p. The key findings are:

- Power-law eigenvalue decay ($\alpha \in [0.63, 1.33]$, $R^2 > 0.93$) is universally preferred over exponential decay.
- Eigenvector delocalization is moderate in high-p datasets and degrades gracefully in low-p settings.
- The spectral prediction σ_K tracks the empirical crud scale with $r > 0.88$ in eight of nine datasets.
- Diagonal concentration holds well for large p and degrades predictably for small p and large K.

- Residual correlations are heavier-tailed than Gaussian, making the error bound in Theorem 3 conservative (the true error is at least as large).
- Cross-component covariances are small in the mean, supporting the negligibility condition.

C Robustness checks

The main text documents approximately power-law eigenvalue spectra across all datasets and argues that the resulting broad spectral structure drives slow decay of the background correlation scale under PC regression. Two natural concerns arise: (i) could the observed approximately power-law spectrum be an artifact of heavy tails, outliers, or scaling properties of the raw data rather than a genuine feature of the dependence structure? and (ii) could the spectral structure be consistent with independence among features, so that the observed spectra are explainable by noise alone? Both concerns can be addressed with systematic robustness checks, and in every case the power-law structure survives.

C.1 Heavy-tail stress test: is the power-law spectrum a preprocessing artifact?

Approximately $1/f$ structure could in principle arise from artifacts such as outlier entries, heterogeneous row scales, or heavy-tailed marginal distributions, rather than from genuine shared dependence among features. If any of these were the dominant driver, then transformations that remove the artifact should noticeably weaken or destroy the power-law behavior. We test this by applying four transformations to each dataset before recomputing the PCA spectrum:

1. **Clipping** at the 99.9th percentile: all entries beyond the 0.1%/99.9% quantiles are clipped to the boundary values, removing extreme outliers while preserving the bulk of the distribution.
2. **Row-wise L_2 normalization**: each row (observation) is divided by its Euclidean norm, removing heterogeneous magnitude scaling across samples.
3. **Rank Gaussianization**: all entries are replaced by their inverse-normal-transformed ranks (via $z = \sqrt{2}\,\mathrm{erfinv}(2u-1)$ where u is the rank-based uniform quantile), completely removing marginal non-Gaussianity while preserving the rank-order dependence structure.
4. **Gaussian null**: an i.i.d. $N(0,1)$ matrix of matching dimensions (capped at 5000×2000) provides a baseline with no dependence structure.

For each transformation, we fit the power law $\log(\lambda_k/\text{total var}) = a - \alpha \log k$ on the first 40 components (matching the procedure in Section B) and report the exponent α and goodness of fit R^2. We also compute the Hill tail index $\hat{\mu}$ on the top 2% of absolute entry values, both on the raw data and after row normalization, to characterize marginal tail heaviness.

The results (Table 8) show that the power-law spectral structure is robust to all transformations. Clipping at the 99.9th percentile has essentially no effect on either α or R^2 in any dataset, ruling out extreme outlier entries as the driver. Row normalization and rank Gaussianization produce modest shifts in α (typically < 0.15) but preserve the qualitative power-law shape with high R^2 (> 0.93 in all cases). The most notable change is that rank Gaussianization can slightly increase α (e.g., RNA-Seq: $1.33 \to 1.44$; Stringer: $0.74 \to 0.89$), suggesting that heavy tails in marginals, if anything, slightly flatten the spectrum rather than create the power law. By contrast, the Gaussian null (i.i.d. noise of matching dimensions) produces a flat spectrum with no power-law structure, confirming

Table 8: Heavy-tail stress test: power-law fits (α, R^2) on the first 40 principal components under different data transformations. The Gaussian null (i.i.d. noise) produces a flat spectrum with no power-law structure.

	Original		Clip@0.999		Row-norm		Rank-Gauss	
Dataset	α	R^2	α	R^2	α	R^2	α	R^2
Kay fMRI	0.67	0.98	0.67	0.98	0.67	0.98	0.67	0.98
Haxby fMRI	1.03	0.98	1.02	0.98	0.87	0.94	0.97	0.96
NHANES	0.63	0.99	0.64	0.99	0.91	0.93	0.83	0.99
CIFAR-10	1.24	1.00	1.24	1.00	1.09	0.99	1.25	1.00
GTEx	1.30	0.99	1.30	0.99	1.24	0.99	1.30	0.99
Precinct	0.70	0.99	0.70	0.99	0.76	0.97	0.77	0.98
RNA-Seq	1.33	0.93	1.33	0.93	1.02	0.94	1.44	0.93
HEXACO	0.97	0.97	0.97	0.97	0.94	0.97	1.02	0.97
Stringer	0.74	0.99	0.74	0.99	0.72	0.99	0.89	0.99

that the observed spectral decay reflects genuine shared dependence rather than a finite-sample artifact.

These checks support treating the broad, approximately $1/f$ eigenvalue spectrum as a real structural property of these datasets. The power-law behavior persists whether or not entries are clipped, rows are rescaled, or marginals are Gaussianized, and it is absent in data with no dependence structure.

C.2 Null-model comparisons: are the spectra consistent with independence?

A complementary question is whether the observed eigenvalue spectra could arise from independent features. If features were truly independent (or had only marginal structure but no cross-feature dependence), then the PCA spectrum should resemble that of a noise matrix. We compare the empirical spectrum of each dataset against two null models:

1. **Gaussian null**: an i.i.d. $N(0, 1)$ matrix of matching dimensions ($n \times p$, capped at 5000×2000). Under the Marchenko–Pastur law [Marčenko and Pastur, 1967], eigenvalues cluster in a bounded interval with no power-law tail.

2. **Column-permutation null**: the values within each column are independently shuffled across rows, destroying all cross-feature correlations while preserving each feature's marginal distribution exactly. This is a stricter control than the Gaussian null because it retains marginal non-Gaussianity, sparsity, and any column-specific scaling. We average over 3 independent permutation realizations.

For each null model, we compute both the explained-variance spectrum and the inverse participation ratio ($\mathrm{IPR}_k = \sum_i v_{ik}^4$) of each eigenvector, which measures eigenvector localization.

Results. The key question is whether the observed spectra could arise from independent features. Across all nine datasets, the empirical eigenvalue spectrum separates dramatically from both null models. The empirical spectra show power-law decay spanning 1–2 orders of magnitude above the null baseline, while both the Gaussian and column-permutation nulls produce flat spectra clustered near $1/p$. The separation is evident from the first principal component onward and persists across

the full range of computed components (up to 300). The broad spectral structure reflects genuine cross-feature dependence rather than marginal properties of individual features.

The column-permutation null is particularly informative: it preserves each feature's exact marginal distribution (including heavy tails, discreteness, and sparsity) but destroys all pairwise and higher-order dependence. The fact that the column-permutation spectrum is indistinguishable from the Gaussian null confirms that the observed power-law structure arises from cross-feature dependence, not from marginal distributional properties.

Eigenvector localization provides a complementary diagnostic. Empirical eigenvectors show moderately elevated IPR values compared to the null models, particularly for the leading components, consistent with the partial eigenvector localization documented in Section B (Test 2). However, the IPR elevation is modest (typically 1.5–3$\times$ the null level) and diminishes for higher-rank components. Eigenvectors are not perfectly delocalized, but their localization is far from extreme enough to invalidate the qualitative conclusions.

C.3 Summary

The robustness checks confirm two key properties of the datasets studied:

- The approximately power-law eigenvalue spectrum is not an artifact of outliers, row-scale heterogeneity, or marginal non-Gaussianity. It persists under clipping, row normalization, and rank Gaussianization, and it is absent in independence-preserving null models (Tables 8).
- The spectral structure reflects genuine cross-feature dependence: column permutation (which preserves marginals but destroys dependence) eliminates the power-law decay entirely.

Together with the assumption audit in Section B, these checks support the main text's interpretation: the broad spectral structure that drives slow decay of the crud scale under PC regression is a real feature of these datasets' dependence structure, not a preprocessing or distributional artifact.

C.4 Alternative adjustment methods: the problem is the spectrum, not PCA

The main text uses top-K PC regression as a deliberately generic adjustment strategy. A natural concern is whether the persistence of broad residual correlations is specific to PCA, or whether it reflects a more fundamental property of the datasets' dependence structure that would limit any moderate-dimensional linear adjustment. We test this by repeating the core analysis—adjust all features, recompute pairwise correlations on residuals, measure the spread—under three alternative adjustment strategies that are commonly used or proposed in observational causal inference practice.

Methods.

1. **Random covariate regression.** For each feature, regress it on a randomly chosen set of m other features (via OLS) and take the residual. This mimics the common practice of "controlling for observables" when no principled adjustment set is available. We average over 5 independent random covariate draws.
2. **Random projection adjustment.** Project the data onto K random Gaussian directions (orthonormalized), regress those out, and measure residual correlation spread. This is a useful control because it tests whether the *choice* of directions matters: if random projections give similar results to PCA, the problem is the dimensionality of shared variation, not the specific directions removed. We average over 5 independent random projection draws.

3. **Lasso adjustment.** For each feature in the correlation subset, use Lasso (ℓ_1-penalized regression, $\alpha = 0.1$) to data-adaptively select predictors from all other features, then compute pairwise correlations on the residuals. This tests whether data-driven variable selection can isolate the "right" confounders and collapse the background dependence. It maps onto doubly-robust and post-double-selection workflows popular in econometrics [Belloni et al., 2014].

For comparability, we measure the same quantity across all methods: the standard deviation of off-diagonal residual correlations on a fixed random subset of 300 features (or all features if $p < 300$), at adjustment dimensions $m \in \{10, 50, 100\}$ for the linear methods.

Table 9: Residual correlation SD under alternative adjustment strategies. PCA targets the top eigenvalues; random covariates mimic observational practice; random projections remove generic directions; Lasso uses data-driven variable selection. All methods leave substantial residual spread.

	Dimension 10			Dimension 50			
Dataset	PCA	Rand. cov.	Rand. proj.	PCA	Rand. cov.	Rand. proj.	Lasso
Kay fMRI	0.036	0.059	0.065	0.027	0.044	0.065	0.032
Stringer	0.019	0.028	0.030	0.015	0.024	0.030	0.017
HEXACO	0.046	0.099	0.145	0.040	0.076	0.146	0.033
Haxby fMRI	0.079	0.129	0.177	0.036	0.075	0.176	0.042
RNA-Seq	0.071	0.125	0.182	0.043	0.067	0.227	0.037
GTEx	0.101	0.170	0.285	0.056	0.095	0.284	0.063
CIFAR-10	0.104	0.139	0.190	0.062	0.090	0.193	0.084
NHANES	0.093	0.097	0.119	0.116	0.105	0.127	0.077
Precinct	0.136	0.135	0.169	0.309†	0.135	0.207	0.094

†Precinct has only p=83 features. Removing 50 of 83 PCs leaves a near-degenerate residual subspace, inflating the residual correlation SD. This is a finite-p artifact, not a property of the adjustment method.

Results (Table 9). The qualitative picture is the same across all methods—all leave substantial residual spread—but the quantitative differences are real and informative.

1. **All methods leave substantial residual spread.** At dimension 10, the residual correlation SD ranges from 0.019 to 0.285 depending on dataset and method; no method collapses it to near zero. Even Lasso, which can draw on all $p - 1$ remaining features as potential predictors (not just K directions), produces residual SDs of 0.017–0.094.

2. **PCA is the most efficient per dimension; random covariates are 1.5–3× worse.** PCA targets the top eigenvalues by construction, so it removes the most shared variance per dimension. Random covariate regression—the procedure closest to common observational practice—produces residual SDs that are typically 1.5–2.2× larger than PCA at the same dimension (e.g., Haxby at dimension 10: PCA 0.079 vs. random covariates 0.129; HEXACO: 0.046 vs. 0.099). This gap reflects the fact that randomly chosen covariates are unlikely to align with the dominant shared directions, so they remove less dependence structure per covariate.

3. **Random projections are nearly inert for high-p datasets.** For CIFAR-10 (p=3072), Haxby (p=1452), and Stringer (p=6000), removing 10, 50, or 100 random directions produces virtually identical residual SDs (e.g., CIFAR-10: 0.190, 0.193, 0.194). This is the cleanest

evidence for the spectral explanation: when the eigenvalue spectrum is broad, random directions miss the shared variance entirely, so projecting them out has negligible effect on the correlation structure.

4. **Lasso can outperform PCA but still cannot collapse the background.** Lasso uses data-driven variable selection from all other features, giving it access to more effective predictors than a fixed K-dimensional subspace. It often produces lower residual SDs than PCA at comparable effective dimensionality (e.g., Haxby: Lasso 0.042 vs. PCA at K=10: 0.079; Precinct: 0.094 vs. 0.136). Nevertheless, even Lasso leaves residual SDs of 0.017–0.094 across datasets, confirming that the background dependence cannot be eliminated by any single-step linear adjustment.

The methods differ by factors of roughly 1.5–3×, which is not negligible—but the central point is that none of them collapse the residual correlation spread to near zero. Whether one uses PCA, random covariates, Lasso, or random projections, residual correlation SDs remain in the range 0.02–0.28. The crud scale is a property of the data's spectral structure, not of the particular adjustment method.

C.5 Additional worked examples

NHANES biomarkers. Hemoglobin and hematocrit are known to be near-deterministically related (hematocrit $\approx 3\times$ hemoglobin by definition). In the NHANES dataset ($n = 29{,}902$), their correlation is $r = 0.91$. Applying the crud-aware z-test with $\sigma_{\text{crud}} = 0.093$ (NHANES at $K = 10$) gives $z_{\text{crud}} = 9.8$, $p_{\text{crud}} < 10^{-20}$. Similarly, ALT and AST (both liver transaminases reflecting hepatocyte damage) correlate at $r = 0.83$, giving $z_{\text{crud}} = 8.9$. Both associations survive the crud-aware calibration by a wide margin, as expected for pairs with known mechanistic relationships.

GTEx RNA-seq. Most cis-eQTLs reported in the GTEx v8 atlas have small effect sizes, with only about 22% showing >2-fold effects [The GTEx Consortium, 2020]. Using the GTEx v8 per-sample TPM data for skeletal muscle (top 10,000 genes by variance across 803 samples), the crud distribution at $K = 10$ has a background SD of 0.10. A typical small cis-eQTL effect of $r = 0.10$ sits at only the 71st crud percentile—squarely within the bulk of background gene–gene correlations. Even $r = 0.20$ reaches only the 95th percentile. Only effects with $r > 0.30$ (the 99th percentile, $3\times$ the background SD) are clearly distinctive. Since most reported cis-eQTLs have small effect sizes, much of the GTEx catalog may not be distinguishable from background transcriptomic dependence using association magnitude alone.

C.6 Positive control: known-strong pairs survive crud-aware calibration

The crud-aware calibration framework (Section A.2) is designed to flag associations that stand out from the background. A natural concern is whether genuinely strong associations—pairs with known biological or psychometric relationships—do in fact survive this calibration. If the crud-aware null swallowed real signal, it would be too conservative to be useful.

We test this using two datasets where ground-truth group structure is available:

- **HEXACO personality inventory** (p=242): items within the same personality facet (e.g., the 10 Honesty-Sincerity items) are designed by psychometricians to measure the same underlying trait. There are 24 facets yielding 1080 within-facet pairs.

- **NHANES biomedical survey** (p=165): features within the same clinical subsystem (e.g., lipid panel: total cholesterol, HDL, triglycerides, LDL; or hematology: RBC, hemoglobin, hematocrit, MCV) are known to be biologically linked. Seven biomarker groups yield 54 within-group pairs.

For each dataset, we compute the residual correlation matrix after removing K PCs, then compute the crud-aware p-value for each known-strong pair: the fraction of all $\binom{p}{2}$ pairs with $|\hat{\rho}^{(K)}| \geq |\hat{\rho}_{ij}^{(K)}|$. If the positive control works, known-strong pairs should be heavily enriched in the extreme tails.

Table 10: Positive control: known-strong pairs are enriched in the tails of the crud-aware null. "Top 5%" and "Top 1%" report the fraction of known-strong pairs with crud-aware p-value below 0.05 and 0.01, respectively (expected: 5% and 1% for random pairs).

Dataset	K	Median $\|r\|$	Median p_{crud}	Top 5%	Top 1%	BG SD
HEXACO	0	0.324	0.041	55%	20%	0.145
HEXACO	10	0.044	0.254	29%	15%	0.046
HEXACO	20	0.048	0.148	33%	14%	0.038
NHANES	0	0.389	0.016	69%	41%	0.115
NHANES	10	0.131	0.085	44%	30%	0.093
NHANES	20	0.390	0.010	70%	54%	0.099

Results (Table 10). The central question is whether the crud-aware null is too aggressive—whether it swallows real signal along with background noise. Known-strong pairs are massively enriched in the tails of the crud-aware null at every level of adjustment, confirming that it does not.

In HEXACO, 55% of within-facet pairs fall in the top 5% at K=0, and 20% fall in the top 1%—enrichments of 11× and 20× over chance. After removing 10 PCs, 29% remain in the top 5% and 15% in the top 1% (6× and 15× enrichment). After 20 PCs, enrichment is similar. The median within-facet $|r|$ drops from 0.32 to 0.04–0.05 after adjustment (comparable to the background SD of 0.04–0.05), yet the enrichment persists because within-facet correlations are systematically larger than typical random-pair correlations at the same adjustment level. Psychometrically designed pairs survive crud-aware calibration at every adjustment level.

In NHANES, the enrichment is even stronger: 69% of biomarker-group pairs are in the top 5% at K=0, rising to 70% at K=20 (where 54% are in the top 1%). This reflects the strong mechanistic links within clinical subsystems (e.g., hemoglobin and hematocrit are physiologically coupled, systolic and diastolic blood pressure are mechanically linked). Biologically grounded pairs are clearly distinctive against the domain background.

The upshot is that the crud-aware calibration framework separates signal from noise in the right direction: genuinely strong associations survive calibration and are clearly distinguished from the background, even after generic PC adjustment. At the same time, the bulk of random-pair associations remain near the crud scale, reinforcing the main text's conclusion that most observed associations in these domains reflect background dependence rather than specific causal links.

References

Michael C. Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166, 2007. doi: 10.1177/1088868306294907.

Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality: An explanation of the $1/f$ noise. *Physical Review Letters*, 59(4):381–384, 1987. doi: 10.1103/PhysRevLett.59.381.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014. doi: 10.1093/restud/rdt044.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020. doi: 10.1111/rssb.12348.

James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001. doi: 10.1126/science.1063736.

Seth J. Hill, Daniel J. Hopkins, and Gregory A. Huber. Local demographic changes and US presidential voting, 2012 to 2016. *Proceedings of the National Academy of Sciences*, 116(50): 25023–25028, 2019. doi: 10.1073/pnas.1909202116.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

Marvin S. Keshner. $1/f$ noise. *Proceedings of the IEEE*, 70(3):212–218, 1982. doi: 10.1109/PROC.1982.12282.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kibeom Lee and Michael C. Ashton. Psychopathy, Machiavellianism, and narcissism in the Five-Factor Model and the HEXACO model of personality structure. *Personality and Individual Differences*, 38(7):1571–1582, 2005. doi: 10.1016/j.paid.2004.09.016.

Yujie Liu, Tingting Geng, Zhenzhen Wan, Qi Lu, Xuena Zhang, Zixin Qiu, Lin Li, Kai Zhu, Liegang Liu, An Pan, and Gang Liu. Associations of serum folate and vitamin B12 levels with cardiovascular disease mortality among patients with type 2 diabetes. *JAMA Network Open*, 5 (1):e2146124, 2022. doi: 10.1001/jamanetworkopen.2021.46124.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Paul E. Meehl. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1):195–244, 1990. doi: 10.2466/pr0.1990.66.1.195.

National Center for Health Statistics. National health and nutrition examination survey (NHANES). `https://www.cdc.gov/nchs/nhanes/`, 2023.

Linda M. O'Keeffe, Gemma Taylor, Rachel R. Huxley, Paul Mitchell, Mark Woodward, and Sanne A. E. Peters. Smoking as a risk factor for lung cancer in women and men: A systematic review and meta-analysis. *BMJ Open*, 8(10):e021611, 2018. doi: 10.1136/bmjopen-2018-021611.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019. doi: 10.1080/07350015.2016.1227711.

William H. Press. Flicker noises in astronomy and elsewhere. *Comments on Astrophysics*, 7: 103–119, 1978.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, 2019. doi: 10.1126/science.aav7893.

Bosiljka Tasic, Zizhen Yao, Lucas T. Graybuck, Kimberly A. Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N. Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018. doi: 10.1038/s41586-018-0654-5.

The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.