High-Dimensional Conditional Independence Testing via Random Projection Aggregation

Dian Jin[1] **Zirui Chen**[2] **Ting Li**[1] **Jiaye Teng**[1*]
[1]School of Statistics and Data Science [2]School of Mathematics
Shanghai University of Finance and Economics

Abstract

Conditional independence (CI) testing is a fundamental problem in statistics. However, many classical nonparametric CI tests suffer from the curse of dimensionality, as data sparsity in high-dimensional spaces degrades local estimation. This paper proposes RP-CIT, a random-projection-aggregated CI test that remains effective in the high-dimensional regime. The central idea is to project X and Y onto random one-dimensional subspaces before applying a univariate base CI test. This projection (i) preserves conditional independence under the null hypothesis and (ii) reduces the original high-dimensional problem to a collection of univariate CI subproblems, thereby avoiding high-dimensional nonparametric estimation. Since a single projection may miss the dependence signal, we aggregate the p-values from k independent projections via a Bonferroni-max rule, which asymptotically controls the Type I error at the nominal level and yields a limiting Type II error bound that decreases exponentially in k under a detectable-projection condition. The framework also supports top-r aggregation for signals spread over several projections and other scalar base tests, such as a projected GCM test for higher-dimensional conditioning variables. Experiments on synthetic benchmarks, image-valued data, known DAGs, and Norman Perturb-seq signatures illustrate its calibration, power, and practical utility. Code is available at `https://github.com/TJ-cyber-kkk/rp-ci`.

1 Introduction

This paper studies conditional independence (CI) testing for continuous random variables with high-dimensional tested variables. Given a triplet $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$, and Z, where X and Y are the tested variables and Z is the conditioning variable, the goal of a CI test is to distinguish between

$$H_0 : X \perp\!\!\!\perp Y \mid Z \qquad \text{versus} \qquad H_1 : X \not\perp\!\!\!\perp Y \mid Z.$$

The default CIT-based RP-CIT procedure analyzed in Section 2 focuses on scalar continuous Z, while Section 3.2 discusses modular variants beyond this setting.

For continuous CI testing, additional distributional or structural assumptions can make the problem substantially more tractable [15, 42]. This paper focuses on nonparametric CI testing. Within this setting, existing CI tests can be roughly grouped into kernel-based [12, 45, 37], regression-based [34, 32], and information-theoretic [31] approaches. While effective in many settings, these methods face dimension-dependent statistical challenges as d_x and d_y grow. Depending on the method, these involve the curse of dimensionality in local nonparametric estimation, dimension-sensitive kernel or neighborhood geometry, or the estimation of multiple nuisance functions and joint calibration of residual-covariance moments. However, modern scientific data increasingly involve high-dimensional vector-valued tested variables in many fields such as genomics [41], making CI testing with high-dimensional tested variables increasingly important.

*Corresponding author: Jiaye Teng (`tengjiaye@sufe.edu.cn`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Algorithm 1 RP-CIT (Random-Projection-Aggregated Conditional Independence Test)

Require: Samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ with $X_i \in \mathbb{R}^{d_x}$, $Y_i \in \mathbb{R}^{d_y}$, $Z_i \in \mathbb{R}$; number of random projections k; number of Monte Carlo reference replicates B; significance level α.
Ensure: Test decision: reject / do not reject $H_0 : X \perp\!\!\!\perp Y \mid Z$.
1: Generate a shared pool of B Monte Carlo reference statistics.
2: **for** $\ell = 1, \ldots, k$ **do**
3: Draw independent random directions $r_x^{(\ell)} \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$ and $r_y^{(\ell)} \sim \mathrm{Unif}(\mathbb{S}^{d_y-1})$.
4: Project the samples: $X_i'^{(\ell)} = \langle r_x^{(\ell)}, X_i \rangle$ and $Y_i'^{(\ell)} = \langle r_y^{(\ell)}, Y_i \rangle$ for $i = 1, \ldots, n$.
5: Apply CIT to $\{(X_i'^{(\ell)}, Y_i'^{(\ell)}, Z_i)\}_{i=1}^n$ using the shared reference pool, and obtain $p_n^{(\ell)}$.
6: **end for**
7: **Reject** H_0 if $\min_{1 \le \ell \le k} p_n^{(\ell)} \le \alpha/k$.

This paper addresses the question of how to test conditional independence when the tested variables are high-dimensional. To address this challenge, this paper proposes a random projection framework that reduces the high-dimensional tested-variable problem to a collection of scalar projected CI subproblems. For each $\ell = 1, \ldots, k$, X and Y are projected onto one-dimensional subspaces,

$$X'^{(\ell)} = \langle r_x^{(\ell)}, X \rangle, \qquad Y'^{(\ell)} = \langle r_y^{(\ell)}, Y \rangle,$$

where $r_x^{(\ell)}$ and $r_y^{(\ell)}$ are independent random directions drawn uniformly from the unit spheres $\mathbb{S}^{d_x-1}$ and $\mathbb{S}^{d_y-1}$, and a univariate CI test is applied to each projected triplet $(X'^{(\ell)}, Y'^{(\ell)}, Z)$.

Since a single random projection may miss the dependence signal due to directional randomness, RP-CIT uses k independent projection pairs and aggregates the resulting p-values via a Bonferroni-max rule [8]. Specifically, let $p_n^{(1)}, \ldots, p_n^{(k)}$ denote the p-values returned by CIT on the k projected subproblems. The Bonferroni-max rule rejects H_0 whenever $\min_{1 \le \ell \le k} p_n^{(\ell)} \le \alpha/k$, that is, whenever at least one projection yields significant evidence against the null after Bonferroni correction. The complete procedure is summarized in Algorithm 1.

In detail, the p-value $p_n^{(\ell)}$ in Line 5 is computed via the distribution-free CIT of Cai et al. [4], whose limiting null distribution depends only on the law of $\mathrm{Unif}(0,1)^3$. A single Monte Carlo reference pool is therefore generated and reused across all k projections to compute the projection-wise p-values.

The Bonferroni-max version of the default CIT-based RP-CIT admits clean theoretical guarantees on both Type I and Type II error. Under regularity conditions inherited from CIT, it asymptotically controls the Type I error at the nominal level for fixed k. Under a detectable-projection condition, projected CIT consistency, and sufficient Monte Carlo resolution, the Type II error has an asymptotic upper bound of $(1-\pi)^k$, where $\pi > 0$ is the probability that a random projection pair falls in a detectable region. This bound decreases exponentially in k. The framework further admits two extensions: top-r aggregation for diffuse alternatives, and an RP-GCM variant that uses the generalized covariance measure (GCM) [34] as the projected base test for richer conditioning variables. The projection leaves Z unchanged, so the difficulty of handling Z remains with the chosen base test.

These theoretical guarantees are supported by experiments on synthetic benchmarks, image-valued CI tasks, known-DAG skeleton recovery, and Norman Perturb-seq data. The synthetic results show that RP-CIT provides stable Type I calibration across diverse null models and has strong power when the dependence signal is captured by a non-negligible set of random projections. Among the compared tests, only RP-CIT combines level control with high power on the nonlinear latent-factor benchmarks at every tested dimension. The image-valued experiment illustrates applicability to tested variables with real-image structure. In known-DAG skeleton recovery with full-rank edge maps, RP-GCM attains the lowest average structural Hamming distance.

The main contributions of this paper are summarized as follows.

- We introduce a null-preserving random-projection aggregation framework for CI testing with high-dimensional tested variables, reducing vector-valued (X, Y) to scalar projected CI subproblems.

- We instantiate the framework as RP-CIT with scalar CIT calibration and Bonferroni-max aggregation, and prove asymptotic Type I control for fixed k and a projection-capture power bound.
- We develop modular extensions, including top-r aggregation and RP-GCM for richer conditioning variables, and validate the approach on synthetic, image-valued, known-DAG, and Perturb-seq skeleton experiments.

2 RP-CIT: Random-Projection-Aggregated CI Test

This section presents RP-CIT, a random-projection aggregation framework for conditional independence testing with high-dimensional tested variables. Section 2.1 introduces the setup, Section 2.2 gives the algorithm and aggregation rule, and Section 2.3 establishes Type I error control and projection-capture power guarantees.

2.1 Notation

Let $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ be i.i.d. samples from (X, Y, Z), where $X \in \mathbb{R}^{d_x}$ and $Y \in \mathbb{R}^{d_y}$ are the tested variables and $Z \in \mathbb{R}$ is the conditioning variable. CI testing is concerned with testing the conditional independence hypothesis

$$H_0 : X \perp\!\!\!\perp Y \mid Z \qquad \text{versus} \qquad H_1 : X \not\perp\!\!\!\perp Y \mid Z.$$

Let $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^d$. For each projection index $\ell = 1, \ldots, k$, draw directions

$$r_x^{(\ell)} \sim \mathrm{Unif}(\mathbb{S}^{d_x-1}), \qquad r_y^{(\ell)} \sim \mathrm{Unif}(\mathbb{S}^{d_y-1}).$$

All $2k$ directions are mutually independent and independent of the observed data. Form the scalar projected variables

$$X_i'^{(\ell)} = \langle r_x^{(\ell)}, X_i \rangle, \qquad Y_i'^{(\ell)} = \langle r_y^{(\ell)}, Y_i \rangle.$$

Let $p_n^{(\ell)}$ denote the p-value returned by the scalar base CI test applied to $\{(X_i'^{(\ell)}, Y_i'^{(\ell)}, Z_i)\}_{i=1}^n$, with the default CIT construction specified in Section 2.2.

2.2 Algorithm

Algorithm 1 summarizes RP-CIT. The procedure has three steps: randomly project the high-dimensional tested variables, apply a scalar base CI test to each projected triplet, and aggregate the resulting projection-wise p-values. We use the distribution-free CIT of Cai et al. [4] as the default base test, while the framework itself is modular.

Base test: CIT. For each projection ℓ, CIT applies the Rosenblatt transform [30] to the scalar triplet $(X'^{(\ell)}, Y'^{(\ell)}, Z)$,

$$U = F_{X'^{(\ell)}|Z}(X'^{(\ell)} \mid Z), \qquad V = F_{Y'^{(\ell)}|Z}(Y'^{(\ell)} \mid Z), \qquad W = F_Z(Z).$$

Under conditional independence, U, V, W are mutually independent $\mathrm{Unif}(0,1)$ variables. CIT computes a dependence statistic $n\widehat{\rho}_n^{(\ell)}$ whose limiting null distribution is distribution-free under the regularity conditions stated below.

For Monte Carlo calibration, independently of the observed data and all projection directions, generate a single pool of $B \geq 1$ independent reference samples, each consisting of n i.i.d. $\mathrm{Unif}(0,1)^3$ draws. Let $n\widehat{\rho}_n^{*(1)}, \ldots, n\widehat{\rho}_n^{*(B)}$ denote the corresponding reference statistics, computed as described in Appendix C.1. The same reference pool is reused across projections. For projection ℓ, we use the plus-one Monte Carlo p-value

$$p_n^{(\ell)} = \frac{1 + \sum_{b=1}^B \mathbf{1}\left\{n\widehat{\rho}_n^{*(b)} \geq n\widehat{\rho}_n^{(\ell)}\right\}}{B+1}.$$

We suppress the dependence on B in the notation.

Bonferroni-max aggregation. Let $p_n^{(1)}, \ldots, p_n^{(k)}$ be the projection-wise p-values. RP-CIT rejects when any projection is significant after Bonferroni correction:

$$\phi_n = \mathbf{1}\left\{ \min_{1 \le \ell \le k} p_n^{(\ell)} \le \alpha/k \right\}.$$

This default rule is effective when one or a few projections capture the conditional dependence signal. When the signal is distributed across several directions, the top-r extension in Section 3.1 aggregates multiple moderate projection-level signals.

Computational cost. Random projection and computation of the projected CIT statistics cost

$$O\{kn(d_x + d_y) + kn^2\},$$

where the first term is random projection and the second is scalar CIT statistic evaluation across k projections. Projection-wise tests are parallelizable. This expression excludes reference-pool construction and preprocessing, as well as p-value lookup. The reference pool is generated once and reused across projections. Runtime diagnostics are reported in Appendix D.3.

2.3 Theoretical properties

This section establishes the theoretical guarantees of RP-CIT. Section 2.3.1 proves Type I error control, while Section 2.3.2 analyzes projection-capture power under alternatives. Unless conditioned otherwise, probabilities in this section are taken over the data, projections, and Monte Carlo randomness.

2.3.1 Type I error control

We first state the projected regularity conditions inherited from the scalar CIT of Cai et al. [4]. Since RP-CIT applies CIT only after projecting X and Y to scalars, these conditions are imposed on a generic projected triplet (X', Y', Z), and are assumed to hold almost surely over the random projection draws.

Assumption 2.1 (Regularity conditions for CIT)**.** The following conditions, due to Cai et al. [4], are imposed on a generic projected triplet (X', Y', Z):

A. Z is a continuous univariate random variable, and for every z in the support of Z, the conditional distribution functions $F_{X'|Z}(\cdot \mid z)$ and $F_{Y'|Z}(\cdot \mid z)$ are continuous.

B. The univariate kernel function $K(\cdot)$ is symmetric about zero, Lipschitz continuous, and satisfies

$$\int K(v)\,dv = 1, \qquad \int v^i K(v)\,dv = 0, \quad 1 \le i \le m-1, \qquad 0 \ne \int v^m K(v)\,dv < \infty.$$

C. The bandwidth h satisfies

$$\frac{nh^2}{\log^2(n)} \to \infty \qquad \text{and} \qquad nh^{4m} \to 0.$$

D. The density $f_Z(z)$ is bounded away from 0 and infinity on its support.

E. The $(m-1)$th derivatives of $F_{X'|Z}(x \mid z) f_Z(z)$, $F_{Y'|Z}(y \mid z) f_Z(z)$, and $f_Z(z)$ with respect to z are locally Lipschitz-continuous.

These are the standard kernel, bandwidth, continuity, and smoothness conditions under which the scalar CIT statistic $n\widehat{\rho}$ converges under the null hypothesis to its distribution-free limit.

Theorem 2.2 (Type I error control of RP-CIT)**.** *Fix $\alpha \in (0,1)$ and $k \ge 1$, and let $(B_n)_{n \ge 1}$ be any deterministic sequence of positive integers. Suppose $P \in H_0$, i.e., $X \perp\!\!\!\perp Y \mid Z$. Assume that, with probability one over $r_x \sim \mathrm{Unif}(\mathbb{S}^{d_x - 1})$ and $r_y \sim \mathrm{Unif}(\mathbb{S}^{d_y - 1})$, the projected triplet*

$$\big(\langle r_x, X\rangle, \langle r_y, Y\rangle, Z\big)$$

satisfies Assumption 2.1. Let $p_n^{(1)}, \ldots, p_n^{(k)}$ be the projection-wise CIT p-values computed using a shared pool of B_n reference statistics, and define

$$\phi_n = \mathbf{1}\left\{ \min_{1 \le \ell \le k} p_n^{(\ell)} \le \alpha/k \right\}.$$

Then

$$\limsup_{n\to\infty} \mathbb{P}_P(\phi_n = 1) \le \alpha.$$

The proof in Appendix E.1 combines null preservation with a distribution-comparison bound for CIT calibration and Bonferroni aggregation. The calibration bound is uniform over the Monte Carlo budget, so B_n need not be fixed.

For growing projection budgets, Appendix E.1 provides a sufficient calibration condition for pointwise asymptotic Type I control: the excess null rejection probability of a single randomly projected test at α/k_n is $o(1/k_n)$. A corresponding rate for a prescribed diverging k_n is not established by the present CIT analysis.

2.3.2 Power analysis

We next analyze power under alternatives. The key question is whether random one-dimensional projections preserve enough conditional dependence signal for aggregation over projected CIT tests to be effective. Let

$$\mathcal{O} := \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$$

be the projection space. For $\omega = (r_x, r_y) \in \mathcal{O}$, define the projected population CIT index

$$\rho(\omega) := \rho(\langle r_x, X\rangle, \langle r_y, Y\rangle \mid Z)\,.$$

By Cai et al. [4, Proposition 1 and Theorem 2], $\rho(\omega) = 0$ if and only if

$$\langle r_x, X\rangle \perp\!\!\!\perp \langle r_y, Y\rangle \mid Z,$$

provided the projected conditional distribution functions are continuous. Thus a projection pair is informative precisely when $\rho(\omega) > 0$.

To ensure that informative projections are not isolated, we impose a smoothness condition whose role is to make $\omega \mapsto \rho(\omega)$ continuous.

Assumption 2.3 (Smooth conditional density conditions)**.** The following conditions are imposed, where $\mathcal{S}_Z := \mathrm{supp}(P_Z)$.

F. The pairs (X, Z) and (Y, Z) admit joint densities $f_{X,Z}(x, z)$ and $f_{Y,Z}(y, z)$ with respect to Lebesgue measure.

G. For Lebesgue-a.e. $x \in \mathbb{R}^{d_x}$, the map $z \mapsto f_{X,Z}(x, z)$ is continuous on $\mathcal{S}_Z$ in the relative topology. Moreover, for every compact set $K \subseteq \mathcal{S}_Z$, there exists an integrable function $M_{X,K} : \mathbb{R}^{d_x} \to [0, \infty)$ such that

$$\sup_{z\in K} f_{X,Z}(x, z) \le M_{X,K}(x) \qquad \text{for a.e. } x.$$

The analogous condition holds for $f_{Y,Z}(y, z)$.

H. The marginal density f_Z is continuous on its support.

Remark 2.4 (Role of Assumption 2.3). Assumption 2.3 is used only for the power analysis. It is a sufficient condition ensuring that the projected population index $\rho(\omega)$ varies continuously with the projection directions, so that one informative projection expands to a positive-measure neighborhood of informative projections. It is not used for Type I error control, and we do not claim it is necessary.

Theorem 2.5 (Power of RP-CIT)**.** *Fix $\alpha \in (0, 1)$ and an integer $k \ge 1$. Let $(B_n)_{n\ge1}$ be a deterministic sequence of positive integers satisfying*

$$\frac{1}{B_n + 1} \le \frac{\alpha}{k} \qquad \textit{for all sufficiently large } n.$$

Suppose $P \in H_1$ satisfies Assumption 2.3. Then there exist a measurable set $N \subseteq \mathcal{O}$ and constants $\epsilon_1, \pi > 0$ such that

$$\rho(\omega) \ge \epsilon_1 \qquad \textit{for all } \omega \in N, \qquad \mathbb{P}_{\mathcal{O}}(\Omega \in N) = \pi > 0.$$

Assume moreover that the k projection pairs are independent and that, for $\mathbb{P}_{\mathcal{O}}$-almost every $\omega \in N$, the projected scalar CIT subproblem satisfies the fixed-alternative consistency conditions of Cai et al. [4, Theorem 5]. Let β_n denote the Type II error of RP-CIT using Bonferroni-max aggregation at level α. Then

$$\limsup_{n\to\infty} \beta_n \le (1 - \pi)^k.$$

The proof is given in Appendix E.2. Here π is the probability that a random projection pair falls in the detectable region N, and $(1-\pi)^k$ is the probability that all k projection pairs miss this region.

For a fixed alternative, the capture analysis also extends to growing projection budgets. Under the projected CIT consistency and Monte Carlo resolution conditions in Corollary E.11, if $\log k_n = o(n)$, then

$$\beta_n \le (1-\pi)^{k_n} + o(1).$$

In particular, $k_n \to \infty$ implies $\beta_n \to 0$.

3 Extensions

This section presents two extensions of RP-CIT, each replacing one component of the framework. Section 3.1 replaces the Bonferroni-max aggregation rule with top-r aggregation, while Section 3.2 replaces the scalar CIT base test with a regression-based test for multivariate conditioning variables Z.

3.1 Beyond max: top-r aggregation

The Bonferroni-max rule rejects only if at least one projection passes the threshold α/k. When the signal is spread over many directions, several projections may carry moderate evidence while none passes α/k. Let $p_{n,(1)} \le \cdots \le p_{n,(k)}$ denote the ordered projection-wise p-values. The top-r rule

$$\phi_n^{(r)}(u) = \mathbf{1}\{p_{n,(r)} \le u\}$$

rejects when at least r projections are significant at level u.

Theorem E.13 uses Markov's inequality to give a valid but conservative threshold $u = \alpha r/k$. A sharper choice is the α-quantile $t_{\alpha,k,r}$ of $\mathrm{Beta}(r, k-r+1)$, the law of $p_{n,(r)}$ for independent $\mathrm{Unif}(0,1)$ p-values. The projected p-values, however, share the sample and the reference pool. Conditionally on this shared randomness $\mathcal{F}_n$, they are i.i.d. with rejection probability $Q_n(u) := \mathbb{P}_P\{p_n(\Omega) \le u \mid \mathcal{F}_n\}$ for a fresh projection Ω, so

$$\mathbb{P}_P\{\phi_n^{(r)}(u) = 1\} = \mathbb{E}_P\big[g_r\{Q_n(u)\}\big], \qquad g_r(q) := \mathbb{P}\{\mathrm{Bin}(k,q) \ge r\}.$$

Because g_r is convex near zero, fluctuations of Q_n raise this probability above $g_r(t_{\alpha,k,r}) = \alpha$. We therefore use the Beta-λ threshold $\tau_\lambda := t_{\alpha,k,r}/\lambda$ with $\lambda > 1$, whose margin $\alpha - g_r(\tau_\lambda) > 0$ must absorb these fluctuations. A second-order expansion of g_r bounds the excess by a multiple of $\mathrm{Var}_P\{Q_n(\tau_\lambda)\}$, which leads to the following condition.

Assumption 3.1 (Conditional left-tail variability)**.** Fix $P \in H_0$, $\alpha \in (0,1)$, $k \ge 2$, $r \in \{2,\dots,k\}$, and $\lambda \ge 1$, and let $L_r := \max_{q\in[0,1]}[g_r''(q)]_+$. Assume

$$\limsup_{n\to\infty} \mathrm{Var}_P\{Q_n(\tau_\lambda)\} \le \frac{2\{\alpha - g_r(\tau_\lambda)\}}{L_r}.$$

Theorem 3.2 (Type I error control of Beta-λ top-r)**.** *If $P \in H_0$ satisfies the conditions of Theorem 2.2 and Assumption 3.1, then* $\limsup_{n\to\infty} \mathbb{P}_P\{\phi_n^{(r)}(\tau_\lambda) = 1\} \le \alpha$.

Assumption 3.1 is sufficient, not a universal calibration of λ: Proposition E.14 gives a checkable version under a Gaussian-residual null, and Appendix D.4 reports empirical diagnostics. The corrected threshold remains much larger than α/k: for $k = 1000$, $\alpha = 0.05$, and $r = 10$, $\alpha/k = 5\times10^{-5}$, $\alpha r/k = 5\times10^{-4}$, and $\tau_{1.5} \approx 3.6\times10^{-3}$. Any gain over Bonferroni-max comes from this larger threshold and is finite-sample: top-r needs r informative projections, so the asymptotic Type II bound $\mathbb{P}\{\mathrm{Bin}(k,\pi) \le r-1\}$ of Theorem E.15 is no smaller than $(1-\pi)^k$. Proofs are given in Appendix E.3.

3.2 Modular extension: RP-GCM for multivariate Z

Random projection reduces X, Y but leaves Z unchanged. As a modular alternative for multivariate Z, RP-GCM replaces the scalar-Z CIT base test with the generalized covariance measure (GCM) [34], using conditional-mean regression instead of conditional CDF estimation.

For projection ℓ, write $X' = X'^{(\ell)}$ and $Y' = Y'^{(\ell)}$, estimate their conditional means given Z by $\widehat{m}_{X'}$ and $\widehat{m}_{Y'}$, and let $R_i = \{X'_i - \widehat{m}_{X'}(Z_i)\}\{Y'_i - \widehat{m}_{Y'}(Z_i)\}$. With $\bar{R}$ and $\widehat{\sigma}^2_R$ the empirical mean and variance of the R_i, the projected p-values

$$p_n^{(\ell)} = 2\{1 - \Phi(|T_n|)\}, \qquad T_n = \sqrt{n}\,\bar{R}/\widehat{\sigma}_R,$$

are aggregated by the Bonferroni-max rule of Section 2.2.

For fixed k, Theorem E.19 shows that RP-GCM controls the Type I error asymptotically at level α under any null distribution for which the GCM regression-error, moment, and nondegeneracy conditions hold for almost every projection pair. Its power targets the residual cross-covariance $\Sigma_{XY|Z}$ between $X - \mathbb{E}(X \mid Z)$ and $Y - \mathbb{E}(Y \mid Z)$. If $\Sigma_{XY|Z} \neq 0$, almost every projection pair captures it, so by Proposition E.20, RP-GCM is consistent for every fixed k whenever the projected tests are consistent. It misses dependence with $\Sigma_{XY|Z} = 0$, such as A5 and A8–A9 in Section 4.1.

Our implementation fits multi-response ordinary least squares (OLS) regressions of X and Y on Z once and projects the in-sample residuals. This is equivalent to fitting OLS after each projection and replaces $2k$ nuisance fits with two. The dimension d_z enters only through these regressions: with linear conditional means, $d_z = o(\sqrt{n})$ suffices asymptotically, whereas finite-sample calibration deteriorates as d_z/n grows. Proofs and diagnostics are in Appendix E.4 and Appendix D.6.

4 Experiments

Section 4.1 evaluates the calibration and power of RP-CIT on synthetic benchmarks, Section 4.2 illustrates its application to CIFAR-10 image patches, and Section 4.3 assesses RP-GCM through known-DAG skeleton recovery and held-out perturbation support on Norman Perturb-seq data. Throughout, we compare against KCI [45], RCIT [37], CMIknn [31], and coordinate-max GCM (Max-GCM) [34], configured as in Appendix C.2.

4.1 Synthetic benchmarks

Setup. We generate $X, Y \in \mathbb{R}^d$ and $Z \in \mathbb{R}$ with $d \in \{10, 50, 200\}$. Unless otherwise stated, $n = 100$, $\alpha = 0.05$, and RP-CIT uses max aggregation over $k = 1000$ projections with a single pool of $B = 500{,}000$ reference statistics. As in the CIT simulations [4], we draw Z from $\mathcal{N}(0, 1)$, whose sparse tails test calibration beyond condition D of Assumption 2.1. Appendix D.2 reports the bounded case, which satisfies this condition: RP-CIT(max) rejects 1.2–2.4% of null replications, and its power changes by at most 4.4 points.

The simulation suite contains four generic nulls N1–N4 with Gaussian, nonlinearly transformed, and heavy-tailed noise, and two types of alternatives. Noise energy grows linearly in d. In the fixed-strength alternatives A1 and A3–A7, signal energy does not grow with d, so noise dilutes the signal. In the scaled alternatives A2, A8, and A9, signal energy grows linearly in d, so the signal-to-noise ratio does not depend on d. A2 spreads a linear signal over all directions. A8 and A9 concentrate a nonlinear signal on one latent factor carried by $\lceil \sqrt{d} \rceil$ coordinates, with $\mathrm{Cov}(X, Y \mid Z = z) = 0$ for every z, unlike A1–A7; their marginal-matched nulls N5 and N6 remove only this dependence. Appendix D.1 specifies all models.

Type I error. Projection restores the calibration of CIT. Direct multivariate CIT [4] rejects all 20,000 null replications for N1–N6 at every d, as Appendix D.2 reports. In Table 1, RP-CIT(max) rejects 1.7–3.8% of replications in every setting, and each null rate varies by at most 0.5 points across d. CMIknn and Max-GCM exceed 5% significantly, reaching 11.2% and 6.7%, and KCI falls to at most 0.2% for $d \geq 50$.

Power. Across Tables 1 and 2, RP-CIT(max) is the only method that combines level control at every d with high power on the nonlinear latent-factor alternatives A8–A9: it attains 70.5–98.9% with no decline in d, whereas the other methods attain at most 42.1% at $d = 10$ and 11.2% for $d \geq 50$. This is consistent with Theorem 2.5: given the alignments of the random directions with the factor loadings, rescaled by $\sqrt{d}$, the projected law does not depend on d, so the probability π of a detectable pair stays away from zero, as Appendix D.1 shows. A2 has the same energy scaling but spreads its signal over all directions, so a random pair retains little of it, and KCI, CMIknn, and Max-GCM lead

Table 1: Empirical Type I error (%) at $\alpha = 0.05$, with $n = 100$, $k = 1000$, and Gaussian Z; each cell pools 10 independent runs of 2,000 replications. N1–N4: linear-Gaussian, quadratic, log–exp, heavy-tailed; N5–N6: marginal-matched nulls of A8–A9. RP: RP-CIT(max); CMI: CMIknn; GCM: Max-GCM. †: significantly above 5% by a one-sided exact binomial test, Holm-adjusted [13] over the table.

	$d = 10$					$d = 50$					$d = 200$				
Model	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM
N1	3.1	0.7	5.3	4.7	5.8†	3.4	0.0	8.5†	4.1	5.9†	3.3	0.0	11.2†	3.7	6.7†
N2	3.3	0.1	4.9	3.4	2.4	3.8	0.0	5.3	3.3	1.3	3.3	0.0	3.1	3.4	0.9
N3	1.7	0.3	1.5	4.4	3.5	2.0	0.0	2.4	4.2	2.2	1.9	0.0	0.1	4.3	1.3
N4	3.0	0.0	2.2	3.1	1.2	3.2	0.0	0.3	3.0	0.6	3.1	0.0	0.2	3.4	0.2
N5	2.2	2.8	4.3	4.5	5.8†	2.6	0.1	5.1	3.9	6.0†	2.4	0.0	5.2	3.8	6.6†
N6	2.1	3.3	4.5	5.3	5.8†	2.4	0.2	5.6†	4.0	6.1†	2.5	0.0	5.2	4.2	6.6†

for $d \geq 50$. RP-CIT(max) attains 95.7% at $d = 50$ and 17.8% at $d = 200$, and Appendix D.2 shows that $n = 500$ restores 100%.

On the fixed-strength alternatives, d affects RP-CIT(max) only through dilution: the scalar test and its calibration do not change with d, but a random pair retains less signal. At $d = 10$, RP-CIT(max) is within one point of the best method on A1, A3, A4, and A6. At $d = 200$, it attains 37.0% on A1, and $n = 500$ restores 100%. On A3–A7, it attains at most 7.7% for $d \geq 50$ and stays near the nominal level at $n = 500$ and $d = 200$. A1 recovers at $n = 500$ because its factor enters both X and Y with sixteen times the noise variance along one direction, whereas A3–A7 couple Y to noise components of X. KCI and Max-GCM lead on A1 at $d = 200$. For $d \geq 50$, Max-GCM leads on A4 and A6, and no method exceeds 15% on A3, A5, and A7.

Table 2: Empirical power (%) at $\alpha = 0.05$, with $n = 100$, $k = 1000$, and Gaussian Z; each cell pools 10 independent runs of 2,000 replications. A1–A7: latent rank-one, dense linear, nonlinear single index, sparse aggregate, product interaction, single-coordinate coupling, Z-modulated; A8–A9: localized square and amplitude. Bold: best entry in each dimension block, with ties within 0.5 points and no bold when all entries are at most 15%.

	$d = 10$					$d = 50$					$d = 200$				
Model	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM
A1	**100**	**100**	**100**	**99.6**	**100**	**100**	**100**	99.2	79.7	**100**	37.0	**91.7**	51.8	18.0	90.8
A2	**100**	**100**	**100**	**99.9**	**100**	95.7	**100**	**100**	75.7	**100**	17.8	**100**	93.9	17.6	62.3
A3	95.5	**96.2**	54.7	47.9	93.5	5.4	0.0	9.6	5.9	12.0	3.3	0.0	6.1	3.8	6.9
A4	**99.9**	77.0	33.9	30.3	**99.7**	7.6	0.0	7.6	4.4	**84.1**	3.5	0.0	5.5	3.7	**46.0**
A5	3.7	2.9	14.3	7.1	5.6	3.3	0.0	6.6	3.9	6.6	3.0	0.0	5.5	3.9	6.8
A6	**99.9**	71.2	48.4	32.7	**100**	7.7	0.0	10.1	4.7	**100**	3.5	0.0	6.4	3.6	**99.7**
A7	66.1	**78.3**	26.7	26.7	66.1	3.9	0.0	7.0	4.4	8.6	3.3	0.0	5.7	3.8	7.0
A8	**70.5**	14.6	37.0	27.5	5.7	**80.9**	3.5	7.3	9.5	6.2	**82.1**	0.1	4.9	4.8	6.7
A9	**96.1**	25.8	42.1	35.7	6.1	**98.5**	4.5	6.6	11.2	6.1	**98.9**	0.1	4.5	5.3	6.7

Sensitivity to k and n. Figure 1 varies k at $d = 50$ for $n \in \{100, 200, 500\}$. Increasing k raises the probability $1 - (1 - \pi)^k$ of capturing an informative direction but tightens the per-projection threshold α/k. Null rejection rates on N1 and N3 stay at or below the nominal 5% level for every tested k at $n = 100$, and Appendix D.2 shows that larger n does not make RP-CIT(max) increasingly conservative. Power increases with k, and larger n makes each informative projection easier to detect, so far fewer projections suffice: A2, whose signal is spread over all directions, exceeds 95% from $k = 1000$ at $n = 100$, $k = 50$ at $n = 200$, and $k = 10$ at $n = 500$, whereas A1 does so from $k = 50$ at $n = 100$.

Top-r aggregation. Appendix D.4.2 applies top-r aggregation with $r \in \{2, 5, 10\}$ and the Beta-λ threshold of Section 3.1, with $\lambda = 1.5$, to the projection p-values behind Tables 1 and 2. It helps where max aggregation has moderate power: at $d = 200$, top-10 raises power from 37.0% to 54.1% on A1 and from 17.8% to 31.4% on A2, and on A8 from 70.5–82.1% to at least 99.7% across d, with

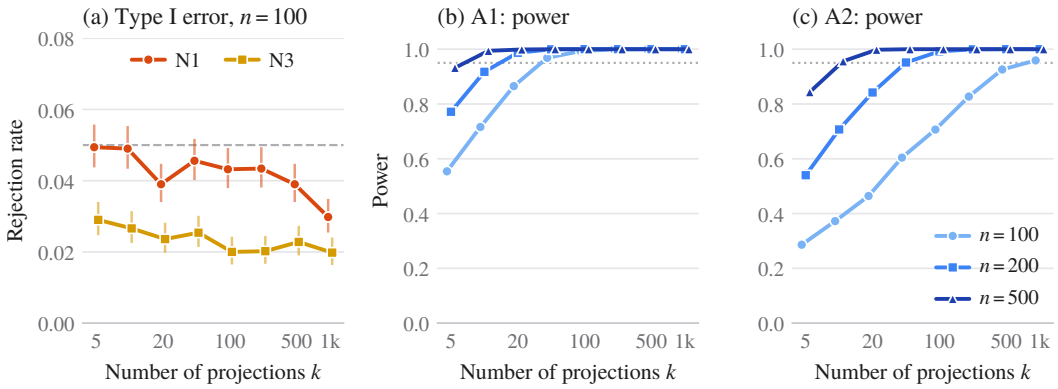

Figure 1: Sensitivity of RP-CIT(max) to k at $d = 50$. (a) Type I error at $n = 100$; (b, c) power on A1 and A2 for $n \in \{100, 200, 500\}$. Dashed and dotted lines mark 5% and 95%. Error bars show pointwise 95% Wilson intervals from 5,000 repetitions at $n = 100$ and 2,000 at $n = 200$ and 500.

no gain on A3–A7 for $d \geq 50$. The only significant null excess is 5.6%, for $r = 10$ on N6 at $d = 10$. We keep max aggregation as the default because its level guarantee needs no condition beyond those of Theorem 2.2, whereas the Beta-λ threshold relies on Assumption 3.1.

4.2 Image-valued CI testing

We next evaluate RP-CIT on real-image structure with $d > n$. We construct a benchmark from CIFAR-10 image patches [17]: two $16 \times 16 \times 3$ patches are vectorized as $X, Y \in \mathbb{R}^{768}$, and a shared scalar covariate Z modulates their brightness. The patches come from the same image with probability η and from two distinct images otherwise, so $\eta = 0$ gives an approximate conditional-independence null, and Appendix D.5 bounds its deviation from an exact null.

RP-CIT(max) controls the level on this benchmark, and its rejection rate generally increases with η and n. With $k = 1000$ and $\alpha = 0.05$, it rejects 3.4–4.8% of replications at $\eta = 0$ for $n \in \{100, 200, 500\}$, and at $\eta = 0.5$ its rejection rate rises from 19.0% to 97.5% as n grows from 100 to 500. Figure 2 in Appendix D.5 compares all five methods: KCI and CMIknn attain higher power in most settings with $\eta > 0$, and Max-GCM rejects 8.5–19.0% at $\eta = 0$.

4.3 Skeleton learning with RP-GCM

With vector-valued nodes, PC-stable [7] conditions on concatenated node vectors, so we use RP-GCM, which handles multivariate Z, as its CI test. Under the standard assumptions [36], CI relations identify a DAG up to its Markov equivalence class, so we evaluate the skeleton, first on known graphs and then against held-out perturbations.

4.3.1 Known-DAG skeleton recovery

We generate data from 30 sparse 10-node DAGs with at most two parents per node, using linear-Gaussian vector SEMs with full-rank edge maps, node dimension $q \in \{10, 30\}$, and sample size $n \in \{500, 1000\}$. All methods share the graphs, datasets, and PC-stable protocol, with $\alpha = 0.05$ and conditioning sets of at most two nodes, under which an exact d-separation oracle recovers all 30 true skeletons.

RP-GCM attains the lowest structural Hamming distance (SHD) in Table 3, less than half that of the next-best CMIknn, and the highest exact recovery rate. Both RP-GCM and Max-GCM recover all true edges, but RP-GCM introduces fewer extra edges: its paired mean SHD reduction is 0.225, with 95% graph-seed-cluster bootstrap interval $[0.133, 0.325]$, and its exact-recovery gain is 0.183, with interval $[0.108, 0.250]$. This gain does not come from conservatism: in matched-null checks, the rejection rates of the two tests differ by at most 1.5 percentage points in every setting, and both over-reject most when d_z/n is large, as Section 3.2 anticipates. With rank-two edge maps, RP-GCM and Max-GCM show a precision–recall trade-off. Appendix D.7 gives complete results and calibration diagnostics.

Table 3: Skeleton recovery for full-rank linear-Gaussian vector SEMs, averaged over four (n, q) settings on the same 30 graph seeds, that is, 120 graph–data settings per method. Table 13 reports matched-null rejection rates.

CI test	SHD ↓	Exact recovery ↑
RP-GCM	0.067	0.933
CMIknn	0.158	0.858
Max-GCM	0.292	0.750
RCIT	1.217	0.442
KCI	9.142	0.000

4.3.2 Norman Perturb-seq

Following Maathuis et al. [23], we test whether signature associations learned on unperturbed cells are supported by held-out perturbations. From a build split of perturbation cells in the Norman Perturb-seq dataset [27], we construct 15 nonoverlapping 30-gene response signatures and learn PC-stable skeletons on 1,500 control cells, conditioning on at most two signatures, so that $d_z \leq 60$. The support of a pair is the larger of the two mean absolute responses of one signature to perturbing the other's target, measured on held-out perturbation cells relative to 582 controls excluded from every graph fit. Source-normalized support uses within-source percentiles to discount broadly strong perturbations. Natural skeletons range from 6 to 105 edges, so each CI test ranks all pairs by its PC edge p-value [18], and we compare its top 10, 30, and 60 pairs.

RP-GCM has the highest held-out support on both measures at 10 pairs and among the CI tests at 30 pairs; at 60 of 105 pairs, this advantage disappears. Signature-mean GCM, correlation, and partial correlation come close to RP-GCM, because held-out support credits indirect effects [6] and so does not separate conditional from marginal association.

RP-GCM also yields a sparse graph whose held-out support exceeds a degree-preserving null. Across ten control-cell subsamples, the edges selected in at least six fits form a 17-edge consensus skeleton, with source-normalized support 0.920 versus 0.865 under this null and nominal $p = 0.0151$. The 68-edge Max-GCM consensus skeleton has 0.736 versus 0.723 and $p = 0.0528$, and the natural 18-edge RP-GCM skeleton has $p = 0.0524$. These edges are perturbation-supported associations, not a ground-truth causal graph; Appendix D.8 gives details.

5 Conclusion

We introduced RP-CIT, a null-preserving random-projection framework for CI testing with high-dimensional tested variables. By reducing vector-valued X and Y to scalar projected subproblems and aggregating calibrated scalar-CIT p-values, RP-CIT avoids direct nonparametric testing in the ambient tested-variable space while retaining asymptotic Type I error control for fixed k. A projection-capture bound relates its power to the probability that a random projection retains detectable dependence. Experiments show stable calibration and, unlike the compared tests, sustained power as dimension grows for nonlinear latent-factor alternatives whose signal energy scales with dimension. Signal dilution lowers power on fixed-strength alternatives, and larger samples restore it when the retained signal is strong enough. The framework further supports modular extensions, including top-r aggregation and RP-GCM for multivariate conditioning variables. RP-GCM attains the lowest average SHD on known DAGs with full-rank edge maps and, on Perturb-seq data, the highest held-out support among the CI tests for its top 10 and 30 pairs.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 6260072784), the Shanghai Science and Technology Development Funds (Grant No. 24YF2711700), the Fundamental Research Funds for the Central Universities (Grant No. 2024110586), and the School-level Project of the School of Statistics and Data Science, Shanghai University of Finance and Economics (Grant No. 2025110083). The authors declare no competing interests.

References

[1] Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, 3rd edition, 2013.

[2] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition, 2003.

[3] Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 2199–2208, 2019.

[4] Zhanrui Cai, Runze Li, and Yaowu Zhang. A distribution free conditional independence test with applications to causal discovery. *Journal of Machine Learning Research*, 23(85):1–41, 2022.

[5] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: "model-X" knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

[6] Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8(1):412, 2025. doi: 10.1038/s42003-025-07764-y.

[7] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014. URL `http://jmlr.org/papers/v15/colombo14a.html`.

[8] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56 (293):52–64, 1961.

[9] Bao Duong and Thin Nguyen. Conditional independence testing via latent representation learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 121–130. IEEE, 2022.

[10] Bao Duong, Nu Hoang, and Thin Nguyen. Amortized conditional independence testing. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 410–423. Springer, 2025.

[11] Jianqing Fan, Yang Feng, and Lucy Xia. A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, 218(1):119–139, 2020.

[12] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, pages 489–496, 2007.

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

[14] Cheng Huang and Xiaoming Huo. A statistically and numerically efficient independence test based on random projections and distance covariance. *Frontiers in Applied Mathematics and Statistics*, 7:779841, 2022.

[15] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014. URL `http://jmlr.org/papers/v15/javanmard14a.html`.

[16] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

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

[18] Junning Li and Z. Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the PC algorithm. *Journal of Machine Learning Research*, 10(17):475–514, 2009. URL `http://jmlr.org/papers/v10/li09a.html`.

[19] Shuangning Li and Molei Liu. Maxway CRT: Improving the robustness of the model-X inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(5):1441–1470, 2023. doi: 10.1093/jrsssb/qkad081.

[20] Xiaotong Lin, Jie Xie, Fangqiao Tian, and Dongming Huang. Testing multivariate conditional independence using exchangeable sampling and sufficient statistics. *arXiv preprint arXiv:2504.06685*, 2025.

[21] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

[22] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[23] Marloes H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010. doi: 10.1038/nmeth0410-247.

[24] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

[25] Ryunosuke Miyazaki and Yoshimasa Uematsu. Testing conditional independence via the spectral generalized covariance measure: Beyond Euclidean data. *arXiv preprint arXiv:2511.15453*, 2025.

[26] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021.

[27] Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

[28] Ivan Nourdin, Giovanni Peccati, and Xiaochuan Yang. Berry-Esseen bounds in the Breuer-Major CLT and Gebelein's inequality. *Electronic Communications in Probability*, 24:1–12, 2019. doi: 10.1214/19-ECP241. Article 34.

[29] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007.

[30] Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23 (3):470–472, 1952.

[31] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947. PMLR, 2018.

[32] Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022.

[33] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

[35] Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 22(285):1–32, 2021.

[36] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.

[37] Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

[38] Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

[39] Liangjun Su and Halbert White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

[40] Mukund Sudarshan, Aahlad Manas Puli, Wesley Tansey, and Rajesh Ranganath. DIET: Conditional independence testing with marginal dependence measures of residual information. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10343–10367. PMLR, 2023.

[41] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

[42] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014. doi: 10.1214/14-AOS1221.

[43] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

[44] Ethan Weinberger, Chris Lin, and Su-In Lee. Isolating salient variations of interest in single-cell data with contrastiveVI. *Nature Methods*, 20(9):1336–1345, 2023. doi: 10.1038/s41592-023-01955-3.

[45] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In Fábio Gagliardi Cozman and Avi Pfeffer, editors, *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 804–813. AUAI Press, 2011. URL `https://webdav.tuebingen.mpg.de/causality/UAI11_KCItest.pdf`.

[46] Yi Zhang, Linjun Huang, Yun Yang, and Xiaofeng Shao. Doubly robust conditional independence testing with generative neural networks. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):195–220, February 2026. doi: 10.1093/jrsssb/qkaf047.

[47] Shunyu Zhao, Yanfeng Yang, Shuai Li, and Kenji Fukumizu. Fast flow matching based conditional independence tests for causal discovery. *arXiv preprint arXiv:2602.08315*, 2026.

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

[49] Li-Ping Zhu, Kai Xu, Runze Li, and Wei Zhong. Projection correlation between two random vectors. *Biometrika*, 104(4):829–843, 2017. doi: 10.1093/biomet/asx043.

A Related work

Assumptions and regimes. Conditional independence (CI) testing depends strongly on the assumptions placed on the joint distribution of (X, Y, Z). For discrete or categorical variables, classical contingency-table procedures and their extensions are available. For continuous data under parametric assumptions, CI often reduces to partial correlation, likelihood-ratio tests, or score tests, with high-dimensional variants based on regularization and debiased inference [1, 2, 24, 15, 42]. Here we focus on fully nonparametric CI testing for continuous data. The default CIT-based RP-CIT further assumes continuous projected variables and a scalar continuous conditioning variable Z, as required by the base test. The random-projection principle itself is more general, but discrete or mixed settings require a compatible scalar base test and calibration.

In the unrestricted nonparametric setting, CI testing is fundamentally difficult: uniformly valid tests over broad null classes cannot achieve nontrivial power without additional assumptions or structure [34, 26]. Useful procedures therefore exploit additional structure or assumptions. Separately, in high-dimensional settings, it is important to distinguish whether dimensionality arises from the conditioning variable Z or from the tested variables X and Y. A major strand of the high-dimensional CI literature focuses on a large conditioning set Z, as in variable selection, model-X inference, and constraint-based causal discovery. Other methods directly accommodate multivariate tested variables X and Y, but d_x and d_y enter method-specific estimation, geometric, learning, or calibration stages. The RP-GCM extension applies the same projection-and-aggregation framework with a different scalar base test for richer Z, while the difficulty of high-dimensional conditioning remains with the chosen base test.

Direct nonparametric CI tests. Kernel-based CI tests characterize conditional independence through RKHS embeddings or conditional cross-covariance operators [12, 45]. KCI is a representative example; scalable variants such as RCIT and RCoT use random Fourier features to reduce computational cost [37, 29]. Related omnibus tests use conditional characteristic functions, weighted Hellinger distance, conditional distance covariance, or conditional distance correlation [38, 39, 43]. For multivariate tested variables, d_x and d_y enter the nonparametric representations or estimators used to construct these test statistics.

Information-theoretic tests characterize conditional dependence through conditional mutual information. Nearest-neighbor methods estimate this quantity using local neighborhoods and can use local permutation for calibration [16, 31]. For multivariate tested variables, d_x and d_y enter the neighborhood geometry, so these local estimates are constructed in the full tested-variable space.

Regression-based procedures avoid full conditional density estimation by estimating conditional mean functions and testing residual covariance or related residual moments. The generalized covariance measure (GCM) and weighted GCM are representative examples [34, 32]. Recent spectral variants approximate conditional cross-covariance structure through data-dependent bases and reduce inference to scalar regressions, extending the GCM idea beyond standard Euclidean data [25]. In coordinate-wise multivariate extensions such as Max-GCM, d_x and d_y enter the nuisance regressions and the number and joint calibration of residual moments.

Taken together, these direct nonparametric CI tests can accommodate multivariate tested variables, but d_x and d_y enter different estimation, geometric, or calibration stages. RP-CIT instead projects X and Y to scalar subproblems before applying the base CI test.

Recent CI testing approaches. Recent CI testing approaches often introduce additional modeling components or alternative testing pipelines, with validity guarantees depending on different assumptions. DIET reformulates CI testing through information residuals such as $F_{X|Z}(X \mid Z)$ and $F_{Y|Z}(Y \mid Z)$, allowing marginal dependence measures to be used for CI testing [40]. LCIT learns latent representations of X and Y that remove information about Z, while amortized approaches such as ACID learn testing rules across synthetic tasks and apply the learned testers to new datasets [9, 10]. CCIT [33] casts CI testing as a classification problem between the joint distribution and a nearest-neighbor approximation to the corresponding conditionally independent reference distribution.

Another line of work constructs or approximates null distributions through conditional resampling. The conditional randomization test (CRT) exploits knowledge of the conditional law $X \mid Z$ to obtain finite-sample Type I error control for arbitrary test statistics [5]. Its extensions consider computational efficiency, robustness to misspecification, and multivariate tested blocks, such as distilled CRT, Maxway CRT, and MS-CRT [21, 19, 20].

Generative CI methods learn conditional distributions or samplers to construct test statistics or calibrate CI tests, including GCIT, DGCIT, MMDCI, and FMCIT [3, 35, 46, 47].

The distribution-free CI test as the base test. The distribution-free CIT of Cai et al. [4] provides the base test used in our default RP-CIT construction. CIT uses Rosenblatt transforms to convert conditional independence into a mutual independence problem among transformed variables with a distribution-free null law. Its limiting null distribution depends only on independent $\mathrm{Unif}(0, 1)$ variables, so critical values can be obtained by simulation. This property makes CIT suitable as the projected-level calibration engine in RP-CIT.

CIT also admits multivariate extensions. However, direct multivariate use requires estimating sequential conditional distribution functions with conditioning sets whose dimensions grow with those of the random vectors. RP-CIT therefore uses CIT after projection: it first reduces the tested variables to scalar subproblems and then applies scalar CIT to each projected triplet. In this construction, CIT provides distribution-free calibration, while random projection provides dimensionality reduction for the tested variables.

Random projections in dependence testing. Random projections have been widely used for high-dimensional dimension reduction and have also appeared in dependence testing. RDC uses random nonlinear projections of empirical copula transformations together with canonical correlation to define an unconditional dependence measure [22]. Projection correlation characterizes unconditional independence of random vectors through projected one-dimensional dependence summaries [49]. RPDC combines random projections with distance covariance to obtain an efficient unconditional multivariate independence test [14]. Projection ideas have also been used in structured conditional settings; for example, Fan et al. [11] develop a projection-based conditional dependence measure under factor-model assumptions for high-dimensional graphical-model learning.

Unlike the unconditional projection-based tests above, RP-CIT uses random projections as a null-preserving reduction to scalar CI subproblems that retain the conditioning variable Z.

Positioning of RP-CIT. RP-CIT provides a modular framework for extending scalar CI tests to vector-valued tested variables. It uses data-independent random projections as a null-preserving reduction before applying a compatible scalar base test, and aggregates projection-wise p-values with

multiplicity control. Shah and Peters [34] suggest a related step for the GCM, applying it to data-independent mappings of (X, Z) and (Y, Z), and leave it to future work. The framework separates dimensional reduction of the tested variables from the choice of scalar CI test: CIT provides the default calibrated inference, while other valid scalar base tests can be incorporated. Its main trade-off is between projection capture and multiplicity, while challenges associated with high-dimensional conditioning remain with the chosen base test.

B Limitations and broader impact

Projection preserves the null for any distribution; the default CIT base test covers continuous projected variables and scalar continuous Z, and other data types need a compatible base test, such as a within-stratum permutation test for finite categorical Z. With CIT, Type I error control is pointwise and asymptotic for fixed k; uniform and growing-k control are left to future work, and choosing k after seeing projection results needs sequential calibration, such as α-spending or e-values. For multivariate Z, RP-GCM inherits the nuisance-regression conditions and residual-covariance target of GCM; our OLS implementation covers linear conditional means.

Power depends on how much signal a random pair retains, which for fixed-strength signals shrinks as d grows. Bonferroni-max aggregation, with per-projection threshold α/k, suits signals that some projection captures strongly. Beta-λ top-r aggregation also uses moderate evidence spread over several projections; its level control needs Assumption 3.1 on the dependence among projected tests, verified in Proposition E.14 for Gaussian residuals in high dimension. Calibrating λ from the data under general dependence and designing other aggregation rules are open directions.

RP-CIT extends CI testing to high-dimensional tested variables in scientific data. Its main risk is reading CI decisions or learned skeletons as causal conclusions outside the stated assumptions; the Perturb-seq analysis therefore reports perturbation-supported associations, not a causal graph.

C Implementation details

C.1 Projected CIT implementation

In RP-CIT, Algorithm 2 uses the shared reference pool generated in Algorithm 1.

For $\gamma \in (0, 1)$, let $q_{n,\gamma}$ denote the order statistic with index

$$B - \lfloor (B+1)\gamma \rfloor + 1$$

in the shared reference pool $\{n\widehat{\rho}_n^{*(b)}\}_{b=1}^{B}$, ordered increasingly, provided $(B+1)\gamma \geq 1$. Otherwise, set $q_{n,\gamma} = +\infty$. For a projection pair ω, write $\widehat{\rho}_n(\omega)$ for its sample CIT index. The plus-one rule then satisfies

$$p_n(\omega) \leq \gamma \quad \Longleftrightarrow \quad n\widehat{\rho}_n(\omega) > q_{n,\gamma}.$$

When $B = B_n$ varies with n, the threshold $q_{n,\gamma}$ is defined using the corresponding reference pool.

C.2 Baseline configurations

KCI and RCIT use causal-learn 0.1.4.3 [48]. KCI uses Gaussian kernels and the Gamma approximation with the median bandwidth, the standard choice for multivariate inputs; the library's default empirical rule gives numerically diagonal kernel matrices for our multivariate unconditional queries. RCIT uses 100 conditioning features and the `lpd4` approximation, with 10 tested-variable features, twice the default, because the tested variables are high-dimensional. CMIknn uses tigramite 5.2.10.1 with 10 nearest neighbors, 5 shuffle neighbors, 100 shuffle samples, and standardized inputs. Max-GCM uses in-sample OLS residuals and 999 Gaussian multiplier draws. These configurations are fixed across all experiments, and every test rejects when $p \leq \alpha$.

Paired sensitivity checks on a separate seed bank support these choices. With 10 tested-variable features, RCIT rejects 42% of alternative and 5% of null replications, versus 32% and 6% with the default 5. With 5 shuffle neighbors, CMIknn rejects 3% of null replications, whereas 10 and 20 raise this rate to 11% and 32%. For KCI and Max-GCM, other null approximations and regressions change rejection rates by at most 0.01.

Algorithm 2 CIT (Distribution-free Conditional Independence Test; [4])

Require: Samples $\{(X_i, Y_i, Z_i)\}_{i=1}^n$; significance level α; kernel $K(\cdot)$ and bandwidth h; number of Monte Carlo draws B; optional precomputed pool $\{n\widehat{\rho}^{*(b)}\}_{b=1}^B$.

Ensure: p-value $\widehat{p}$ and test decision for $H_0 : X \perp\!\!\!\perp Y \mid Z$.

Step 1: Rosenblatt transform via nonparametric CDF estimation.

Define $K_h(t) = K(t/h)/h$ and $\widehat{f}_Z(z) = \frac{1}{n}\sum_{j=1}^n K_h(z - Z_j)$.

for $i = 1, \ldots, n$ **do**

$$\widehat{U}_i \leftarrow \widehat{F}_{X|Z}(X_i \mid Z_i) = \frac{\frac{1}{n}\sum_{j=1}^n K_h(Z_i - Z_j)\mathbf{1}(X_j \le X_i)}{\widehat{f}_Z(Z_i)}.$$

$$\widehat{V}_i \leftarrow \widehat{F}_{Y|Z}(Y_i \mid Z_i) = \frac{\frac{1}{n}\sum_{j=1}^n K_h(Z_i - Z_j)\mathbf{1}(Y_j \le Y_i)}{\widehat{f}_Z(Z_i)}.$$

$$\widehat{W}_i \leftarrow \widehat{F}_Z(Z_i) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}(Z_j \le Z_i).$$

end for

Step 2: Compute the CIT statistic $\widehat{\rho}$.

Let $c_0 = (13e^{-3} - 40e^{-2} + 13e^{-1})^{-1}$.

Compute

$$\widehat{\rho} = c_0 n^{-2} \sum_{i=1}^n \sum_{j=1}^n \Big(e^{-|\widehat{U}_i - \widehat{U}_j|} + e^{-\widehat{U}_i} + e^{\widehat{U}_i - 1} + e^{-\widehat{U}_j} + e^{\widehat{U}_j - 1} + 2e^{-1} - 4\Big) \times \Big(e^{-|\widehat{V}_i - \widehat{V}_j|} + e^{-\widehat{V}_i} + e^{\widehat{V}_i - 1} + e^{-\widehat{V}_j} + e^{\widehat{V}_j - 1} + 2e^{-1} - 4\Big) e^{-|\widehat{W}_i - \widehat{W}_j|}.$$

Step 3: Generate or reuse the Monte Carlo reference pool.

if no reference pool is supplied **then**

for $b = 1, \ldots, B$ **do**

Independently across b and of the observed sample, draw i.i.d. $(U_i^{*(b)}, V_i^{*(b)}, W_i^{*(b)}) \sim \mathrm{Unif}(0,1)^3$ for $i = 1, \ldots, n$.

Compute $\widehat{\rho}^{*(b)}$ by the formula in Step 2 with $(\widehat{U}_i, \widehat{V}_i, \widehat{W}_i)$ replaced by $(U_i^{*(b)}, V_i^{*(b)}, W_i^{*(b)})$.

end for

end if

Step 4: Plus-one p-value and decision.

Compute

$$\widehat{p} = \frac{1 + \sum_{b=1}^B \mathbf{1}\{n\widehat{\rho}^{*(b)} \ge n\widehat{\rho}\}}{B + 1}.$$

Return $\widehat{p}$; reject H_0 if $\widehat{p} \le \alpha$, and do not reject otherwise.

D Additional experimental details

D.1 Synthetic data-generating mechanisms

The formulas below match the released generators.

Common setup. Observations (X_i, Y_i, Z_i), $i = 1, \ldots, n$, are i.i.d. with $X_i, Y_i \in \mathbb{R}^d$ and scalar Z_i. In the Gaussian suite $Z_i \sim \mathcal{N}(0,1)$; in the bounded suite $Z_i \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$, which has unit variance and satisfies Assumption 2.1(D). Both suites use the same formulas. Unless stated otherwise, the noise vectors $E_{x,i}, E_{y,i}$ have i.i.d. $\mathcal{N}(0,1)$ coordinates, and scalar functions act coordinate-wise. A random unit vector is $g/\|g\|_2$ with $g \sim \mathcal{N}(0, I_d)$. Coefficients, directions, and index sets are drawn once per replication and held fixed across its n observations; all noise and latent variables are mutually independent and independent of Z.

Generic nulls (N1–N4). $X_i = Z_i \mathbf{1}_d + f(E_{x,i})$ and $Y_i = Z_i \mathbf{1}_d + g(E_{y,i})$, with

N1 (linear-Gaussian): $f(e) = g(e) = e$;

N2 (quadratic): $f(e) = e^2 - 1, \quad g(e) = e$;

N3 (log–exp): $f(e) = \log(1 + |e|), \quad g(e) = \exp\{\mathrm{clip}(e, -2, 2)\} - 1$;

N4 (heavy-tailed): $f(e) = \sin e, \quad g(e) = e^2 - 1, \quad$ with i.i.d. t_5 coordinates of $E_{x,i}, E_{y,i}$.

Alternatives A1–A7. Let $\beta_x, \beta_y \in \mathbb{R}^d$ have i.i.d. $\mathcal{N}(0, 0.5^2)$ entries, and set $X_{0,i} = Z_i\beta_x + \frac{1}{2}E_{x,i}$ and $Y_{0,i} = Z_i\beta_y + \frac{1}{2}E_{y,i}$. Unless modified below, $X = X_0$ and $Y = Y_0$. Let u, v be independent random unit vectors, $H_i \sim \mathcal{N}(0,1)$, $c \sim \mathcal{N}(0,1)$, and $\sigma(t) = 1/(1 + e^{-t})$; subscripts index coordinates.

A1 (latent rank-one): $X_i = X_{0,i} + \gamma H_i u, \quad Y_i = Y_{0,i} + \theta H_i v$;

A2 (dense linear): $Y_i = Y_{0,i} + \theta M X_{0,i}, \quad M_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$;

A3 (nonlinear single index): $Y_i = Y_{0,i} + \theta \tanh(u^\top X_{0,i})\, v$;

A4 (sparse aggregate): $Y_{i1} = Y_{0,i1} + \theta\, |S|^{-1/2} \sum_{j \in S} X_{0,ij}, \quad S$ a random 5-subset;

A5 (product interaction): $Y_{i1} = Y_{0,i1} + \theta X_{0,i1} X_{0,i2}$;

A6 (single-coordinate coupling): $Y_{ij} = Y_{0,ij} + \theta X_{0,ij}, \quad j = \lfloor d/2 \rfloor + 1$;

A7 (Z-modulated): $Y_i = Y_{0,i} + \theta\, \sigma(cZ_i)\, (u^\top X_{0,i})\, v$.

In Section 4.1, $\theta = \gamma = 2$. Under A5, $\mathrm{Cov}(X, Y \mid Z = z)$ is linear in z and averages to zero, so $\Sigma_{XY|Z} = 0$. Noise energy grows linearly in d. Signal energy grows at the same rate on A2, where its expectation is $\theta^2 d/4$, and does not grow with d on the other alternatives.

Localized alternatives (A8–A9) and matched nulls (N5–N6). Let $\theta = 1.5$, $\kappa = \theta/\sqrt{1 + \theta^2}$, and $s = \lceil \sqrt{d} \rceil$. Draw β_x, β_y as above, $S_i, T_i \sim \mathcal{N}(0,1)$, and unit vectors u, v whose supports are independent uniformly random s-subsets, with i.i.d. $\mathcal{N}(0,1)$ weights on the support before normalization. Set

$$X_i = Z_i\beta_x + \tfrac{1}{2}\big(\kappa\sqrt{d}\, S_i\, u + \sqrt{1 - \kappa^2}\, E_{x,i}\big),$$
$$Y_i = Z_i\beta_y + \tfrac{1}{2}\big(\kappa\sqrt{d}\, f(W_i)\, v + \sqrt{1 - \kappa^2}\, E_{y,i}\big),$$

with $W_i = S_i$ under the alternatives and $W_i = T_i$ under the matched nulls, which therefore keep the conditional marginals of X and Y and remove only the shared factor. A8 and N5 use $f(t) = (t^2 - 1)/\sqrt{2}$; A9 and N6 use $f(t) = (|t| - \sqrt{2/\pi})/\sqrt{1 - 2/\pi}$. Both functions are even and standardized, so $\mathbb{E}\{Sf(S)\} = 0$: under A8–A9, X and Y have zero conditional cross-covariance given Z but are dependent through S. The factor $\sqrt{d}$ makes signal energy grow linearly in d. Let $a = \sqrt{d}\,\langle r_x, u\rangle$ and $b = \sqrt{d}\,\langle r_y, v\rangle$, which have unit second moment at every d. By the rotation invariance of the Gaussian noise and of β_x, β_y, the law of the projected triplet given (a, b) does not depend on d. Since (a, b) converges in distribution to a pair of independent standard Gaussians, the probability π of Theorem 2.5 stays away from zero as d grows.

D.2 Additional synthetic results

Tables 4 and 5 repeat Tables 1 and 2 with bounded Z.

Direct multivariate CIT applies CIT to (X, Y, Z) without projection, with the same bandwidth rule as RP-CIT and reference pools simulated at the matching dimensions. It rejects all 20,000 null replications for N1–N6 at every $d \in \{10, 50, 200\}$ under both Gaussian and bounded Z, so its power is not reported. To locate the breakdown, we apply it to N1 with Gaussian Z, $d_x = d_y \in \{1, 2, 5\}$, and $n \in \{100, 500\}$, using 200 replications per setting and $B = 10^4$ reference statistics for $d_x = 2, 5$. Its rejection rates are 7.0% and 5.5% for $d_x = 1$, 5.5% and 2.5% for $d_x = 2$, and 100% at both n for $d_x = 5$: its calibration breaks down between $d_x = 2$ and 5, and a larger sample does not restore it.

Larger sample sizes. Table 6 reports RP-CIT(max) at $d = 200$ and $k = 1000$ for $n \in \{100, 200, 500\}$; runs with $n = 200$ and 500 use reference pools of $B = 10^6$ statistics simulated at the matching n. No null rejection rate at $n = 200$ or 500 is significantly above 5% by a

Table 4: Empirical Type I error (%) at $\alpha = 0.05$, with $n = 100$, $k = 1000$, and bounded $Z \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$; each cell pools 10 independent runs of 2,000 replications. N1–N4: linear-Gaussian, quadratic, log–exp, heavy-tailed; N5–N6: marginal-matched nulls of A8–A9. RP: RP-CIT(max); CMI: CMIknn; GCM: Max-GCM. †: significantly above 5% by a one-sided exact binomial test, Holm-adjusted over the table.

	$d = 10$					$d = 50$					$d = 200$				
Model	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM
N1	2.1	0.6	4.8	5.2	5.7†	2.4	0.0	5.1	4.9	5.8†	2.1	0.0	5.7†	4.7	6.8†
N2	2.2	0.1	4.9	3.6	2.3	2.4	0.0	5.0	4.1	1.2	2.4	0.0	5.9†	4.3	0.9
N3	1.2	0.2	1.2	4.6	3.4	1.4	0.0	0.2	4.9	2.0	1.4	0.0	0.0	5.1	1.3
N4	2.3	0.0	2.1	3.4	1.2	2.2	0.0	0.1	3.4	0.7	2.4	0.0	0.1	4.2	0.3
N5	1.6	2.6	4.5	5.0	5.6†	1.8	0.1	4.5	4.4	6.2†	1.8	0.0	4.8	4.4	6.6†
N6	1.5	3.1	4.5	5.5	5.8†	1.6	0.2	4.9	4.7	6.0†	1.8	0.0	4.8	4.8	6.4†

Table 5: Empirical power (%) at $\alpha = 0.05$, with $n = 100$, $k = 1000$, and bounded $Z \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}]$; each cell pools 10 independent runs of 2,000 replications. A1–A7: latent rank-one, dense linear, nonlinear single index, sparse aggregate, product interaction, single-coordinate coupling, Z-modulated; A8–A9: localized square and amplitude. Bold: best entry in each dimension block, with ties within 0.5 points and no bold when all entries are at most 15%.

	$d = 10$					$d = 50$					$d = 200$				
Model	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM	RP	KCI	CMI	RCIT	GCM
A1	**100**	**100**	**100**	**99.7**	**100**	**100**	**100**	98.9	80.5	**100**	34.7	**92.2**	48.5	18.8	91.1
A2	**100**	**100**	**100**	**99.9**	**100**	96.1	**100**	**100**	76.9	**100**	14.8	**100**	92.1	19.1	62.1
A3	93.7	**94.7**	51.5	47.6	93.0	4.2	0.0	7.9	6.7	12.4	2.3	0.0	5.4	4.9	7.3
A4	**99.9**	76.9	32.8	31.2	**99.8**	5.9	0.0	6.3	5.5	**83.8**	2.6	0.0	5.2	4.7	**45.6**
A5	2.4	4.0	**17.4**	9.9	5.8	2.2	0.0	6.4	4.6	5.9	2.2	0.0	5.1	4.5	6.6
A6	**99.9**	71.1	47.9	33.0	**100**	6.2	0.0	8.9	5.4	**100**	2.5	0.0	5.5	4.4	**99.7**
A7	64.0	**78.6**	25.5	27.8	66.5	2.8	0.0	6.3	5.5	8.1	2.3	0.0	5.0	4.5	6.8
A8	**74.9**	14.2	37.1	27.3	5.3	**85.2**	3.4	7.0	9.7	6.0	**85.9**	0.1	4.3	5.5	6.6
A9	**97.4**	24.7	42.0	35.3	6.1	**99.2**	3.8	6.0	12.0	6.2	**99.5**	0.1	4.2	5.9	6.9

one-sided exact binomial test, unadjusted, and every null rate at $n = 500$ is closer to 5% than at $n = 100$.

Table 6: RP-CIT(max) rejection rates (%) at $d = 200$, $k = 1000$, $\alpha = 0.05$, and Gaussian Z. The $n = 100$ column repeats Tables 1 and 2; for $n = 200$ and 500, each null uses 2,000 replications and each alternative 1,000.

Model	$n = 100$	200	500	Model	$n = 100$	200	500
N1	3.3	2.3	4.3	A1	37.0	90.4	100
N2	3.3	3.9	4.1	A2	17.8	52.3	100
N3	1.9	1.7	3.2	A3	3.3	4.3	4.9
N4	3.1	2.7	3.6	A4	3.5	2.4	5.5
N5	2.4	2.3	3.7	A5	3.0	2.2	4.8
N6	2.5	2.5	3.4	A6	3.5	2.2	5.7
				A7	3.3	3.4	4.6
				A8	82.1	100	100
				A9	98.9	100	100

D.3 Runtime and scalability

Setup. We time RP-CIT on the linear-Gaussian null N1 with scalar Z and $d_x = d_y = d$, using $\alpha = 0.05$, bandwidth scale 0.8, $B = 10^6$ references, and Bonferroni-max aggregation. Each test runs on a single CPU core, with one BLAS thread and no projection-level parallelization. Timings include projection, the k projected CIT statistics, and p-value computation and aggregation, but exclude data generation and the one-time construction of the reference pool.

Results. Runtime is linear in k and nearly constant in d, which enters only through the projection step. It grows superlinearly in n because of the pairwise computations in the scalar CIT statistic.

Table 7: Single-core wall-clock runtime of RP-CIT. Each entry is the median of 10 runs after one warm-up run; interquartile ranges are within 4% of the medians.

Experiment	Fixed setting	Varied value	Runtime (s)
Vary k	$n = 100, d = 50$	$k = 100$	0.031
	$n = 100, d = 50$	$k = 500$	0.139
	$n = 100, d = 50$	$k = 1000$	0.276
	$n = 100, d = 50$	$k = 2000$	0.549
Vary n	$k = 1000, d = 50$	$n = 100$	0.276
	$k = 1000, d = 50$	$n = 200$	1.676
	$k = 1000, d = 50$	$n = 500$	17.205
Vary d	$n = 100, k = 1000$	$d = 50$	0.277
	$n = 100, k = 1000$	$d = 200$	0.283
	$n = 100, k = 1000$	$d = 500$	0.297

D.4 Additional details for top-r aggregation

This subsection checks Assumption 3.1 empirically and reports the finite-sample behavior of top-r aggregation. Throughout, $n = 100$, $k = 1000$, $\alpha = 0.05$, $\lambda = 1.5$, and Z is Gaussian.

D.4.1 Empirical diagnostic for the variability condition

Assumption 3.1 bounds $\mathrm{Var}_P\{Q_n(\tau_\lambda)\}$ by the margin $m_r := 2\{\alpha - g_r(\tau_\lambda)\}/L_r$. For each null model N1–N6 and $d \in \{10, 50, 200\}$, we generate 50 datasets and, for each, estimate $Q_n(\tau_\lambda)$ by the fraction $\widehat{Q}$ of $M = 10^5$ fresh projections whose p-value is at most τ_λ, using the reference pool of Table 1. We take the sample variance of $\widehat{Q}$ across datasets, subtract the projection-sampling term $\widehat{Q}(1-\widehat{Q})/M$, and add the reference-pool term s_B of Proposition E.14, which is below 1% of m_r for every r. Table 8 reports the ratio of the resulting variance estimate to m_r; a ratio below one is consistent with Assumption 3.1. For N1–N4, all point estimates are below one except for N1 at $d = 10$ and $r = 10$, at 1.03; upper bootstrap limits exceed one only at $d = 10$, and every ratio is at most 0.27 for $d \geq 50$. The marginal-matched nulls N5–N6 keep a latent factor within each of X and Y. Their ratios exceed one at $d = 10$ for every r, at 4.5–8.7 on N5 and 23–31 on N6, and in 8 of the 12 cells with $d \geq 50$; only N5 at $d = 10$ with $r = 10$ has a lower bootstrap limit above one. The assumption is sufficient rather than necessary.

D.4.2 Finite-sample results

Tables 9 and 10 compare max, Beta-λ, and Markov top-r aggregation on the replications of Tables 1 and 2. All rules are applied to the same projection p-values in each replication, so differences between columns reflect only the aggregation rule, and the max column equals those tables.

Beta-λ rejects at most 2.7% of null replications under N1–N4 and at most 5.6% under N5–N6. The latter rate, for $r = 10$ on N6 at $d = 10$, is the only one significantly above 5%; Table 8 gives the largest ratio of the variability condition, 30.6, in that setting. The Markov rule rejects at most 0.9%. For every r, Beta-λ improves on max in all settings where max has power between 15% and 99%: A1 and A2 at $d = 200$, A2 at $d = 50$, A3 and A7 at $d = 10$, and A8–A9 at every d. Where max is near the nominal level, top-r does not help. The Markov rule improves on max by more than one point only on A8–A9 and is otherwise often far less powerful; on A1 at $d = 200$ it attains 21.3%, 4.2%, and 0.3% for $r = 2, 5, 10$, against 37.0% for max. It needs no condition beyond marginal validity, but its threshold $\alpha r/k$ is two to seven times smaller than τ_λ.

Table 8: Estimated $\mathrm{Var}_P\{Q_n(\tau_\lambda)\}$ divided by the margin m_r of Assumption 3.1, with 95% bootstrap intervals over 50 datasets. Values below one are consistent with the assumption.

Model	d	$r = 2$	$r = 5$	$r = 10$
m_r		5.2×10^{-8}	8.6×10^{-7}	2.7×10^{-6}
N1	10	0.71 [0.34, 1.08]	0.72 [0.39, 1.01]	1.03 [0.61, 1.42]
	50	0.09 [0.04, 0.15]	0.12 [0.05, 0.20]	0.19 [0.09, 0.32]
	200	0.03 [0.01, 0.05]	0.05 [0.03, 0.07]	0.12 [0.07, 0.17]
N2	10	0.34 [0.15, 0.53]	0.59 [0.26, 0.97]	0.96 [0.48, 1.46]
	50	0.05 [0.02, 0.07]	0.06 [0.03, 0.09]	0.10 [0.06, 0.13]
	200	0.04 [0.01, 0.07]	0.06 [0.04, 0.09]	0.12 [0.07, 0.18]
N3	10	0.24 [0.11, 0.41]	0.33 [0.16, 0.58]	0.56 [0.28, 0.96]
	50	0.06 [0.02, 0.11]	0.08 [0.03, 0.15]	0.15 [0.05, 0.26]
	200	0.06 [0.02, 0.09]	0.07 [0.04, 0.10]	0.13 [0.07, 0.18]
N4	10	0.41 [0.22, 0.59]	0.52 [0.30, 0.72]	0.88 [0.56, 1.18]
	50	0.12 [0.05, 0.19]	0.16 [0.07, 0.26]	0.27 [0.12, 0.43]
	200	0.05 [0.01, 0.12]	0.05 [0.02, 0.10]	0.09 [0.03, 0.17]
N5	10	4.54 [0.11, 11.42]	5.61 [0.46, 13.66]	8.66 [1.21, 20.44]
	50	0.21 [0.05, 0.49]	0.51 [0.11, 1.18]	1.04 [0.25, 2.32]
	200	2.19 [0.04, 6.18]	2.43 [0.09, 6.72]	3.69 [0.21, 9.97]
N6	10	24.24 [0.12, 68.99]	23.49 [0.37, 65.06]	30.56 [0.99, 82.64]
	50	1.96 [0.07, 5.46]	2.85 [0.20, 7.64]	4.64 [0.50, 11.85]
	200	0.19 [0.08, 0.29]	0.44 [0.17, 0.77]	1.08 [0.38, 1.92]

Table 9: Empirical Type I error (%) of max, Beta-λ, and Markov top-r aggregation on the replications of Table 1. Thresholds: α/k for max, $t_{\alpha,k,r}/1.5$ for Beta-λ, and $\alpha r/k$ for Markov. †: significantly above 5% by a one-sided exact binomial test, Holm-adjusted over the table.

			Beta-λ			Markov		
Model	d	max	$r = 2$	$r = 5$	$r = 10$	$r = 2$	$r = 5$	$r = 10$
Threshold ($\times 10^{-3}$)		0.05	0.24	1.31	3.62	0.10	0.25	0.50
N1	10	3.1	2.4	2.5	2.2	0.6	0.0	0.0
	50	3.4	1.7	1.0	0.5	0.3	0.0	0.0
	200	3.3	1.7	1.0	0.4	0.3	0.0	0.0
N2	10	3.3	2.4	2.7	2.3	0.6	0.0	0.0
	50	3.8	2.0	1.2	0.6	0.4	0.0	0.0
	200	3.3	1.9	1.1	0.5	0.4	0.0	0.0
N3	10	1.7	0.8	0.4	0.2	0.2	0.0	0.0
	50	2.0	0.7	0.1	0.0	0.1	0.0	0.0
	200	1.9	0.6	0.1	0.1	0.1	0.0	0.0
N4	10	3.0	2.3	2.2	1.8	0.5	0.0	0.0
	50	3.2	1.7	0.9	0.4	0.3	0.0	0.0
	200	3.1	1.7	0.9	0.3	0.2	0.0	0.0
N5	10	2.2	2.2	4.0	4.9	0.7	0.4	0.2
	50	2.6	1.9	2.8	3.2	0.6	0.1	0.1
	200	2.4	1.6	2.3	2.7	0.4	0.1	0.1
N6	10	2.1	2.5	4.3	5.6†	0.9	0.5	0.3
	50	2.4	2.0	3.2	4.0	0.6	0.3	0.1
	200	2.5	2.2	3.2	3.8	0.6	0.3	0.1

Table 10: Empirical power (%) of max, Beta-λ, and Markov top-r aggregation on the replications of Table 2. Thresholds are as in Table 9.

			Beta-λ			Markov		
Model	d	max	$r=2$	$r=5$	$r=10$	$r=2$	$r=5$	$r=10$
Threshold ($\times 10^{-3}$)		0.05	0.24	1.31	3.62	0.10	0.25	0.50
A1	10	100	100	100	100	100	100	100
	50	100	100	100	100	100	100	100
	200	37.0	42.8	54.4	54.1	21.3	4.2	0.3
A2	10	100	100	100	100	100	100	100
	50	95.7	99.6	100	100	95.9	92.0	75.2
	200	17.8	20.3	30.1	31.4	5.9	0.3	0.0
A3	10	95.5	97.4	98.7	98.8	94.3	90.0	81.6
	50	5.4	3.6	3.0	1.8	0.8	0.0	0.0
	200	3.3	1.9	1.0	0.4	0.3	0.0	0.0
A4	10	99.9	100	100	100	99.8	99.6	98.6
	50	7.6	5.8	5.6	3.6	1.4	0.0	0.0
	200	3.5	1.9	1.2	0.4	0.3	0.0	0.0
A5	10	3.7	2.8	3.4	3.4	0.7	0.0	0.0
	50	3.3	1.7	1.0	0.5	0.3	0.0	0.0
	200	3.0	1.6	0.9	0.4	0.3	0.0	0.0
A6	10	99.9	100	100	100	99.9	99.5	98.6
	50	7.7	6.1	5.5	3.9	1.4	0.0	0.0
	200	3.5	1.8	1.0	0.4	0.4	0.0	0.0
A7	10	66.1	75.9	85.3	86.5	56.4	36.1	16.7
	50	3.9	2.5	1.6	0.8	0.5	0.0	0.0
	200	3.3	1.6	1.1	0.4	0.2	0.0	0.0
A8	10	70.5	91.0	98.8	99.7	75.6	78.9	77.6
	50	80.9	95.6	99.7	99.9	84.1	85.2	82.1
	200	82.1	96.1	99.6	99.9	84.9	85.9	82.7
A9	10	96.1	99.7	100	100	97.8	98.5	98.4
	50	98.5	99.9	100	100	99.2	99.5	99.3
	200	98.9	99.9	100	100	99.2	99.4	99.3

D.5 Image-valued CI testing: construction and comparisons

Construction. We use a fixed pool of 60,000 CIFAR-10 images (training and test combined), with pixels scaled to $[0, 1]$. For each observation, select source image a uniformly and b uniformly subject to $b \neq a$. Independently draw $Z \sim \mathrm{Unif}[-1, 1]$ and $C \sim \mathrm{Bernoulli}(\eta)$. Apply brightness scaling $I \mapsto \mathrm{clip}\{(1 + 0.4Z)I, 0, 1\}$, independent pixelwise $N(0, 0.02^2)$ noise, and clipping to $[0, 1]$. The upper-left $16 \times 16 \times 3$ patch of a forms X. The lower-right patch of a if $C = 1$, or of b otherwise, forms Y. The noises for X and Y are independent. Both patches are vectorized to 768 dimensions without learned image features or further standardization. Because $b \neq a$ is enforced, $\eta = 0$ is not an exact conditional-independence null; drawing a and b independently gives one. The $\eta = 0$ law is this null law conditioned on $a \neq b$, an event of probability $1 - 1/60{,}000$ per observation, so the two n-sample laws, and hence the rejection probabilities of any test, differ by at most $n/60{,}000 \leq 0.0083$.

Methods. We vary $n \in \{100, 200, 500\}$ and $\eta \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$ at $\alpha = 0.05$. RP-CIT(max) uses $k = 1000$, bandwidth scale 0.8, and $B = 10^6$ shared references. All five methods share the first 200 data seeds; each baseline uses 200 repetitions per setting, and RP-CIT uses 1,000 at $\eta = 0$ and 200 otherwise. Max-GCM uses all 768^2 coordinate pairs; the remaining baseline configurations are given in Appendix C.2.

Baseline comparison. Figure 2 reports all five methods across the 18 settings.

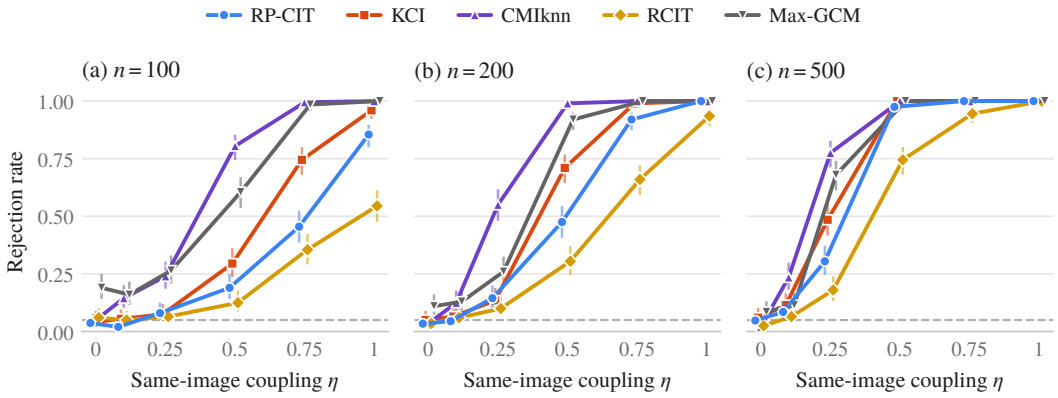

Figure 2: CIFAR-10 baseline comparison at $\alpha = 0.05$. Panels correspond to $n = 100, 200, 500$. Error bars show pointwise 95% Wilson intervals.

D.6 RP-GCM diagnostics

We compare RP-GCM with Max-GCM on identical datasets using the shared in-sample OLS residuals of Appendix E.4.4. RP-GCM uses $k = 1000$ projections, normal p-values, and Bonferroni aggregation; Max-GCM uses 999 Gaussian multiplier draws with plus-one calibration. We fix $d_x = d_y = 10$ and $\alpha = 0.05$, with 2,000 paired repetitions per setting.

Let $Z \sim \mathcal{N}(0, I_{d_z})$, and let $b(Z) \in \mathbb{R}^{10}$ be the first ten coordinates of Z, repeated cyclically when $d_z < 10$. With independent $\varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, I_{10})$, set $X_0 = b(Z) + \varepsilon_X$ and $Y_0 = b(Z) + \varepsilon_Y$. The models are

N1: $X = X_0, \quad Y = Y_0;$

N2: $X = b(Z) + \varepsilon_X^{\odot 2} - \mathbf{1}, \quad Y = Y_0;$

A1: $X = X_0 + 0.5Hu, \quad Y = Y_0 + 0.5Hv, \quad H \sim \mathcal{N}(0, 1);$

A2: $X = X_0, \quad Y = Y_0 + 0.3MX_0, \quad M_{ab} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/10);$

A7: $X = X_0, \quad Y = Y_0 + 0.8\,\sigma(c^\top Z)\,(u^\top X_0)\,v, \quad c \sim \mathcal{N}(0, I_{d_z}),$

where u, v are random unit vectors and M, u, v, c are drawn once per replication. Both nulls have linear conditional means, so OLS is correctly specified under the null. The residual cross-covariances are $0.25\,uv^\top$ (A1), $0.3\,M^\top$ (A2), and $0.4\,uv^\top$ (A7).

Table 11 reports two scans. At $n = 2000$, null rejection rates stay near 0.05 for every d_z. RP-GCM is more powerful than Max-GCM on A1 (0.996–0.997 versus 0.931–0.940), whose rank-one residual cross-covariance spreads the signal over all coordinate pairs (Appendix E.4.3). On A2 and A7, both methods reject in all 2,000 replications for every d_z (not tabulated). At $d_z = 50$, null rejection rates exceed 0.05 for small n, most clearly for RP-GCM (0.111 on N1 and 0.086 on N2 at $n = 500$, versus 0.090 and 0.054 for Max-GCM), and approach 0.05 as n grows. This matches the degrees-of-freedom effect of Appendix E.4.4: the variance factor $\{1 - (d_z + 1)/n\}^{-1}$ is 1.11, 1.05, 1.03, and 1.01 for $n = 500$, 1000, 2000, and 5000. RP-GCM, whose per-projection threshold α/k lies further in the tail, is affected more.

Table 11: Rejection rates of RP-GCM and Max-GCM from 2,000 paired repetitions per setting. Panel (a) varies d_z at $n = 2000$. Panel (b) varies n at $d_z = 50$.

(a) Conditioning dimension d_z

Model	Method	1	5	10	50
N1	RP-GCM	0.041	0.053	0.050	0.056
	Max-GCM	0.050	0.049	0.046	0.052
N2	RP-GCM	0.046	0.041	0.046	0.041
	Max-GCM	0.042	0.043	0.049	0.050
A1	RP-GCM	0.996	0.996	0.996	0.997
	Max-GCM	0.931	0.931	0.940	0.938

(b) Sample size n

Model	Method	500	1000	2000	5000
N1	RP-GCM	0.111	0.065	0.056	0.055
	Max-GCM	0.090	0.067	0.052	0.049
N2	RP-GCM	0.086	0.064	0.041	0.037
	Max-GCM	0.054	0.048	0.050	0.040

D.7 Known-DAG skeleton recovery

Data. We generate 30 sparse 10-node DAGs using a random topological order, edge probability 0.24, and at most two parents per node, retaining graphs with 5–12 edges. Each node lies in $\mathbb{R}^q$, with $q \in \{10, 30\}$, and data follow the linear-Gaussian vector SEM

$$M_j = \sum_{i \in \mathrm{pa}(j)} W_{ij}^\top M_i + \varepsilon_j, \qquad \varepsilon_j \sim N(0, I_q).$$

Full-rank edge maps combine Gaussian-QR orthogonal factors with singular values in $[0.21, 0.6]$, the largest equal to 0.6; rank-two maps are rescaled products of two $q \times 2$ Gaussian matrices with spectral norm 0.6. For each graph and q, the $n = 500$ sample is nested within the $n = 1000$ sample. Baseline configurations are given in Appendix C.2.

Skeleton recovery. With full-rank edge maps, RP-GCM, Max-GCM, and KCI miss no true edge, so their SHD in Table 12 counts extra edges only; RCIT errs mostly by missing edges. With rank-two edge maps, the paired SHD reduction of RP-GCM relative to Max-GCM is 0.175 (95% CI $[-0.092, 0.442]$); at $n = 1000, q = 30$, RP-GCM has higher precision (0.995 versus 0.888) but lower recall (0.716 versus 0.873). Paired intervals, here and in Section 4.3.1, use 20,000 bootstrap resamples of the 30 graph seeds, averaging the four (n, q) contrasts within each seed.

Table 12: Known-DAG recovery over 30 matched graphs per setting. Entries are mean SHD, counting extra and missing edges, with the exact recovery rate in parentheses.

n	q	RP-GCM	Max-GCM	CMIknn	RCIT	KCI
Full-rank edge maps						
500	10	0.100 (0.900)	0.333 (0.767)	0.100 (0.900)	0.467 (0.600)	8.833 (0.000)
500	30	0.033 (0.967)	0.200 (0.833)	0.333 (0.700)	3.100 (0.067)	9.100 (0.000)
1000	10	0.067 (0.933)	0.400 (0.633)	0.133 (0.900)	0.200 (0.800)	8.033 (0.000)
1000	30	0.067 (0.933)	0.233 (0.767)	0.067 (0.933)	1.100 (0.300)	10.600 (0.000)
Rank-two edge maps						
500	10	0.000 (1.000)	0.767 (0.467)	5.167 (0.000)	5.067 (0.000)	2.933 (0.033)
500	30	6.900 (0.000)	6.567 (0.000)	8.000 (0.000)	8.633 (0.000)	1.567 (0.167)
1000	10	0.033 (0.967)	0.667 (0.500)	2.700 (0.033)	2.567 (0.067)	3.567 (0.000)
1000	30	2.500 (0.000)	2.133 (0.133)	7.833 (0.000)	8.367 (0.000)	2.200 (0.067)

Matched-null diagnostics. To assess calibration in the regime of the PC queries, we generate conditional-independence data matching the sample size, node dimension, and conditioning depth

$s \in \{0, 1, 2\}$. For $s > 0$, let $Z \sim N(0, I_{sq})$ and

$$X = B_X^\top Z + \varepsilon_X, \qquad Y = B_Y^\top Z + \varepsilon_Y,$$

where the coefficient entries are independent $N\{0, 0.45^2/(sq)\}$ draws and the two noise vectors are independent $N(0, I_q)$. For $s = 0$, X and Y are independent standard Gaussian vectors. Table 13 reports the rejection rates.

Table 13: Matched-null rejection rates at nominal $\alpha = 0.05$. s is the conditioning depth and $d_z = sq$. RP-GCM, Max-GCM, CMIknn, and RCIT use 1,000 repetitions per setting (95% Wilson half-widths at most 0.031); KCI uses 100.

n	q	s	d_z	RP-GCM	Max-GCM	CMIknn	RCIT	KCI
500	10	0	0	0.048	0.049	0.053	0.040	0.04
		1	10	0.043	0.050	0.061	0.253	0.95
		2	20	0.045	0.059	0.047	0.309	0.99
	30	0	0	0.047	0.046	0.049	0.052	0.04
		1	30	0.086	0.076	0.099	0.401	1.00
		2	60	0.134	0.135	0.070	0.549	1.00
1000	10	0	0	0.053	0.065	0.050	0.038	0.05
		1	10	0.049	0.055	0.065	0.143	0.73
		2	20	0.057	0.042	0.058	0.162	0.98
	30	0	0	0.057	0.059	0.050	0.049	0.06
		1	30	0.060	0.057	0.120	0.194	1.00
		2	60	0.079	0.081	0.089	0.472	1.00

D.8 Norman Perturb-seq: signatures and held-out support

Signatures and data separation. We use the processed subset of the K562 CRISPRa data of Norman et al. [27] distributed with contrastiveVI [44], which contains 27,658 cells, including 7,275 negative-control-guide cells, and 2,000 highly variable genes. Expression is $\log(1 + x)$, where x is the UMI count scaled so that each cell's total over these genes equals the median total across cells. Eligible single-perturbation cells are split into build, orientation-diagnostic, and evaluation sets of 40%, 30%, and 30%; the orientation-diagnostic set is not used here. Among targets with at least 100 eligible cells, we select 15 using mean absolute build-split effects relative to a fixed mean of 1,500 control cells. Candidate genes exclude RPL*, RPS*, MT-*, MALAT1, NEAT1, and the target's own gene. All eligible target–gene pairs are sorted globally by absolute build effect, and greedy assignment selects 30 genes per target, with no gene assigned twice. These are target-specific response signatures, not predefined pathways. The signatures and cell splits are recorded in `01_SIGNATURES_EXACT.json` and `split_cells.json`.

Graph fitting and baselines. We learn skeletons on 1,500 control cells with PC-stable at $\alpha = 0.05$, conditioning on at most two signature nodes, so that $d_z \leq 60$. At each depth, PC tests every conditioning set drawn from the union of the two endpoints' current adjacencies. RP-GCM uses $k = 1000$ and in-sample OLS, and Max-GCM uses the same regression fit. OLS is a first-order model of the conditional means: if they are nonlinear, the OLS residuals retain a function of Z, and Remark E.21 no longer guarantees calibration. RP-GCM admits other regression learners that satisfy Assumption E.18. We also evaluate KCI, RCIT, and CMIknn, configured as in Appendix C.2, and Mean-GCM, the GCM test applied to signature means. Correlation and partial correlation of signature means provide rankings without PC.

Independent held-out support. Evaluation perturbation cells enter neither signature selection, graph fitting, nor pair ranking. For target a and gene g, let $\Delta_{a,g}$ be the held-out perturbation mean minus the mean of 582 reference controls excluded from every main or subsample graph fit. For target b's 30-gene signature $\mathcal{G}_b$, define

$$s_{a\to b} = \frac{1}{30} \sum_{g \in \mathcal{G}_b} |\Delta_{a,g}|, \qquad s_{ab} = \max(s_{a\to b}, s_{b\to a}).$$

Source-normalized support replaces each $s_{a\to b}$ by its rank among the 14 other targets of the same source a, divided by 14, with average ranks for ties, and then takes the larger direction. Edge-set support is the mean score over the selected undirected pairs.

Matched candidate-pair budgets. Natural PC skeletons vary in size: RP-GCM, Max-GCM, Mean-GCM, KCI, RCIT, and CMIknn return 18, 73, 65, 105, 6, and 78 edges, respectively. KCI rejects all 9,660 CI tests of PC-stable and returns the complete graph, in line with its liberal matched-null calibration in Table 13. For size-matched comparisons, each PC-based method ranks all 105 pairs in ascending order of its edge p-value [18], the largest p-value among its *tested* conditioning sets, and its natural skeleton therefore consists of the pairs with edge p-value at most α. The tested sets depend on each method's PC path. Ties, as at the Max-GCM minimum p-value of $1/1000$ or the RP-GCM cap of 1, are broken by the weakest test statistic among the sets attaining that p-value, in descending order, then by the smallest tested p-value and pair name. The correlation baselines rank pairs by absolute correlation or partial correlation of signature means. In Table 14, RP-GCM has the highest support on both measures at 10 pairs and, among the PC-based tests, at 30 pairs. At every budget up to 54 pairs, not only those tabulated, it exceeds Max-GCM on both measures; at 60 pairs, KCI and Max-GCM exceed it on both measures.

Table 14: Mean held-out support of the top 10, 30, and 60 pairs of each method on the Norman data, relative to the 582 reference controls. The upper panel uses raw support and the lower panel source-normalized support.

Method	10 pairs	30 pairs	60 pairs
Raw support			
RP-GCM	0.3579	0.3132	0.2671
Max-GCM	0.2437	0.2616	0.2772
Mean-GCM	0.3186	0.2868	0.2654
KCI	0.2568	0.2623	0.2778
RCIT	0.2398	0.2793	0.2656
CMIknn	0.2152	0.2217	0.2421
Mean correlation	0.3344	0.2966	0.2697
Mean partial corr.	0.3140	0.2982	0.2764
Source-normalized support			
RP-GCM	0.9500	0.8214	0.7560
Max-GCM	0.7500	0.7571	0.7607
Mean-GCM	0.8857	0.8048	0.7440
KCI	0.7929	0.7643	0.7762
RCIT	0.8143	0.7571	0.7393
CMIknn	0.6500	0.6738	0.6952
Mean correlation	0.9000	0.8310	0.7560
Mean partial corr.	0.9000	0.8214	0.7595

Subsample consensus. Ten overlapping 1,500-control-cell subsamples yield consensus skeletons by retaining edges selected in at least six fits. RP-GCM and Max-GCM yield 17 and 68 edges, respectively, with mean edge counts of 17.1 ± 2.0 and 70.3 ± 3.0 across subsamples. Against 10,000 degree-preserving edge-swap reference graphs, RP-GCM has source-normalized support of 0.9202 versus a reference mean of 0.8646, with nominal $p = 0.0151$; for Max-GCM, the values are 0.7363 and 0.7232, with $p = 0.0528$. The natural 18-edge RP-GCM skeleton has nominal $p = 0.0524$.

Interpretation. A retained edge is an undirected association between two signatures that no tested set of at most two other signatures separates; it does not indicate direct or directed regulation. Among the three RP-GCM consensus pairs with the highest held-out support, selected before consulting annotations, only CEBPA–SPI1 shares a gene-set enrichment, MYC targets with adjusted $p = 0.0082$, which gives context rather than validation.

E Proofs and additional theoretical results

E.1 Proof of Theorem 2.2

The proof combines the null-preserving property of random projections (Lemma E.1) with a distribution-comparison bound for CIT calibration (Lemma E.3). The resulting marginal calibration bound also yields the growing-k condition in (2).

Lemma E.1 (Random projection preserves conditional independence)**.** *If $X \perp\!\!\!\perp Y \mid Z$, then for every fixed $r_x \in \mathbb{S}^{d_x-1}$ and $r_y \in \mathbb{S}^{d_y-1}$,*

$$\langle r_x, X\rangle \perp\!\!\!\perp \langle r_y, Y\rangle \mid Z.$$

Proof. Define $f\colon \mathbb{R}^{d_x} \to \mathbb{R}$ by $f(x) = \langle r_x, x\rangle$ and $g\colon \mathbb{R}^{d_y} \to \mathbb{R}$ by $g(y) = \langle r_y, y\rangle$. Both maps are Borel measurable. Let $A, B \subseteq \mathbb{R}$ be arbitrary Borel sets. Since $X \perp\!\!\!\perp Y \mid Z$, for any Borel sets $\widetilde{A} \subseteq \mathbb{R}^{d_x}$ and $\widetilde{B} \subseteq \mathbb{R}^{d_y}$,

$$\mathbb{P}\big(X \in \widetilde{A},\ Y \in \widetilde{B} \mid Z\big) \;=\; \mathbb{P}\big(X \in \widetilde{A} \mid Z\big)\,\mathbb{P}\big(Y \in \widetilde{B} \mid Z\big) \quad \text{a.s.}$$

Taking $\widetilde{A} = f^{-1}(A)$ and $\widetilde{B} = g^{-1}(B)$, and noting that $\{X \in f^{-1}(A)\} = \{f(X) \in A\}$ and $\{Y \in g^{-1}(B)\} = \{g(Y) \in B\}$, we obtain

$$\mathbb{P}\big(f(X) \in A,\ g(Y) \in B \mid Z\big) \;=\; \mathbb{P}\big(f(X) \in A \mid Z\big)\,\mathbb{P}\big(g(Y) \in B \mid Z\big) \quad \text{a.s.}$$

Since A and B are arbitrary, this is precisely

$$\langle r_x, X\rangle \perp\!\!\!\perp \langle r_y, Y\rangle \mid Z.$$

□

Definition E.2 (Asymptotically valid p-value)**.** Let $\{p_n\}_{n\geq 1}$ be a sequence of $[0,1]$-valued statistics for testing a null hypothesis H_0. We say that $\{p_n\}_{n\geq 1}$ is *asymptotically valid* if, for every distribution $P \in H_0$ and every $u \in (0,1)$,

$$\limsup_{n\to\infty} \mathbb{P}_P(p_n \leq u) \leq u.$$

Lemma E.3 (Asymptotic validity of CIT)**.** *Let X', Y' be univariate and let Z be continuous. Suppose Assumption 2.1 holds for (X', Y', Z), and $X' \perp\!\!\!\perp Y' \mid Z$. For each integer $B \geq 1$, let $p_{n,B}$ be the plus-one CIT p-value computed from B i.i.d. reference statistics, independent of the observed sample, as in Algorithm 2. Then*

$$\sup_{B\geq 1} \sup_{u\in(0,1)} \big[\mathbb{P}_P(p_{n,B} \leq u) - u\big]_+ \longrightarrow 0.$$

In particular, the p-values are asymptotically valid in the sense of Definition E.2 for any deterministic sequence $B = B_n \geq 1$.

Proof. Let F_n and G_n be the distribution functions of $n\widehat{\rho}_n$ and a single independent reference statistic $n\widehat{\rho}_n^*$, respectively.

Under the null hypothesis and Assumption 2.1, Cai et al. [4, Theorems 3 and 4] give

$$n\widehat{\rho}_n \xrightarrow{d} T_\infty, \qquad n\widehat{\rho}_n^* \xrightarrow{d} T_\infty,$$

where

$$T_\infty = c_0 \sum_{j=1}^{\infty} \lambda_j \chi_j^2(1),$$

and the $\chi_j^2(1)$ are independent. The limiting kernel is nonzero and positive semidefinite, so at least one eigenvalue is strictly positive. The corresponding scaled chi-square variable has a continuous distribution and is independent of the remaining sum. Hence T_∞ has a continuous distribution function F_∞.

Since F_∞ is continuous, both distribution functions converge uniformly to F_∞, and hence

$$\sup_{t\in\mathbb{R}} |F_n(t) - G_n(t)| \to 0.$$

Fix $B \geq 1$ and $u \in (0,1)$, and let $m = \lfloor (B+1)u \rfloor$. If $m = 0$, the plus-one p-value cannot be at most u, so the rejection probability is zero.

Otherwise, let $q_{n,u}$ be the reference order statistic defined in Appendix C.1. By independence of the observed statistic and the reference pool,

$$\begin{aligned}\mathbb{P}_P(p_{n,B} \leq u) &= \mathbb{E}[1 - F_n(q_{n,u})] \\ &\leq \mathbb{E}[1 - G_n(q_{n,u})] + \sup_t |F_n(t) - G_n(t)| \\ &\leq \frac{\lfloor (B+1)u \rfloor}{B+1} + \sup_t |F_n(t) - G_n(t)|.\end{aligned}$$

For an independent observation with distribution G_n, exchangeability gives a super-uniform plus-one p-value. Evaluating it at $m/(B+1)$ yields the last inequality, including in the presence of ties.

Since the distributional discrepancy does not depend on B or u,

$$\sup_{B \geq 1} \sup_{u \in (0,1)} [\mathbb{P}_P(p_{n,B} \leq u) - u]_+ \leq \sup_t |F_n(t) - G_n(t)| \longrightarrow 0.$$

□

Proof of Theorem 2.2. Fix $P \in H_0$. Let Ω denote an independent random projection pair with distribution $\mathbb{P}_{\mathcal{O}}$.

By Lemma E.1, every fixed projection preserves conditional independence. By assumption, the projected triplet satisfies Assumption 2.1 for $\mathbb{P}_{\mathcal{O}}$-almost every ω.

Let $F_{n,P}$ be the distribution function of the observed statistic $n\widehat{\rho}_n(\Omega)$, including randomness from the data and projection:

$$F_{n,P}(t) = \int_{\mathcal{O}} \mathbb{P}_P(n\widehat{\rho}_n(\omega) \leq t) \; d\mathbb{P}_{\mathcal{O}}(\omega).$$

For almost every fixed ω, the scalar CIT null limit gives

$$n\widehat{\rho}_n(\omega) \xrightarrow{d} T_\infty.$$

Since F_∞ is continuous, bounded convergence over the projection distribution yields

$$F_{n,P}(t) \to F_\infty(t) \qquad \text{for every } t \in \mathbb{R}.$$

This convergence is uniform in t because F_∞ is continuous.

Let G_n be the reference distribution function from the proof of Lemma E.3, and define

$$d_n(P) := \sup_{t \in \mathbb{R}} |F_{n,P}(t) - G_n(t)|.$$

Since G_n also converges uniformly to F_∞, we have

$$d_n(P) \longrightarrow 0.$$

Applying the distribution-comparison argument from Lemma E.3 to a single random projection gives, for every $B \geq 1$ and $u \in (0,1)$,

$$\mathbb{P}_P\{p_{n,B}(\Omega) \leq u\} \leq \frac{\lfloor (B+1)u \rfloor}{B+1} + d_n(P) \leq u + d_n(P).$$

Now fix $k \geq 1$ and let $B = B_n$ be any deterministic sequence of positive integers. All projection-wise p-values have the same marginal distribution, even though they share the observed data and reference pool. The Bonferroni union bound therefore gives

$$\begin{aligned}\mathbb{P}_P(\phi_n = 1) &\leq \sum_{\ell=1}^{k} \mathbb{P}_P\Big(p_n^{(\ell)} \leq \frac{\alpha}{k}\Big) \\ &\leq \alpha + k\, d_n(P).\end{aligned}$$

Since k is fixed and $d_n(P) \to 0$,

$$\limsup_{n \to \infty} \mathbb{P}_P(\phi_n = 1) \leq \alpha.$$

□

Calibration inheritance for growing projection budgets. For a single randomly projected CIT p-value, define its excess null rejection probability by

$$\Delta_{n,B}(u;P) := \left[\mathbb{P}_P\{p_{n,B}(\Omega) \le u\} - u\right]_+, \qquad u \in (0,1).$$

The probability includes the observed data, projection direction, and independent Monte Carlo reference pool.

For RP-CIT with projection budget k and reference budget B, the same union bound gives

$$\mathbb{P}_P(\phi_n = 1) \le \alpha + k\Delta_{n,B}(\alpha/k;P). \tag{1}$$

The distribution comparison above also yields $\Delta_{n,B}(u;P) \le d_n(P)$.

Consequently, for deterministic sequences $k = k_n$ and $B = B_n$, a sufficient condition for pointwise asymptotic Type I control is

$$k_n\Delta_{n,B_n}(\alpha/k_n;P) \to 0. \tag{2}$$

For a class $\mathcal{P}_0 \subseteq H_0$, define

$$\Delta_{n,B}(u;\mathcal{P}_0) := \sup_{P\in\mathcal{P}_0} \Delta_{n,B}(u;P).$$

Then

$$\sup_{P\in\mathcal{P}_0} \mathbb{P}_P(\phi_n = 1) \le \alpha + k_n\Delta_{n,B_n}(\alpha/k_n;\mathcal{P}_0).$$

Thus the corresponding class-level condition

$$k_n\Delta_{n,B_n}(\alpha/k_n;\mathcal{P}_0) \to 0$$

is sufficient for uniform asymptotic level control.

Although $d_n(P) \to 0$ under the stated CIT regularity conditions for each fixed null distribution P, these pointwise limits do not supply a calibration rate for a prescribed diverging k_n, nor do they imply uniform convergence over $\mathcal{P}_0$.

E.2 Proofs for the power analysis

This appendix proves Theorem 2.5 and its growing-k extension (Corollary E.11). We first establish a positive-probability region of detectable projections (Corollary E.9) using an informative projection and continuity of the projected CIT index under Assumption 2.3. We then combine projection capture, average estimation error, and a Monte Carlo reference tail bound (Lemma E.10) to derive both results.

E.2.1 Existence of an informative projection

Lemma E.4 (Existence of an informative projection)**.** *If $X \not\perp\!\!\!\perp Y \mid Z$, then there exist directions $(r_x^\star, r_y^\star) \in \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$ such that*

$$\langle r_x^\star, X\rangle \not\perp\!\!\!\perp \langle r_y^\star, Y\rangle \mid Z.$$

Under the projected CIT conditions ensuring that the population index is well defined and characterizes conditional independence, this is equivalent to

$$\rho(r_x^\star, r_y^\star) > 0.$$

Proof. We argue by contradiction. Suppose that for every $(r_x, r_y) \in \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$,

$$\langle r_x, X\rangle \perp\!\!\!\perp \langle r_y, Y\rangle \mid Z. \tag{3}$$

Let $\mathcal{D}_x \subset \mathbb{S}^{d_x-1}$ and $\mathcal{D}_y \subset \mathbb{S}^{d_y-1}$ be countable dense subsets, and let $\mathbb{Q}$ denote the rationals. For each $(r_x, r_y, s_1, s_2) \in \mathcal{D}_x \times \mathcal{D}_y \times \mathbb{Q} \times \mathbb{Q}$, (3) implies

$$\mathbb{E}\Big[e^{i(s_1 r_x^\top X + s_2 r_y^\top Y)} \,\Big|\, Z\Big] = \mathbb{E}\Big[e^{is_1 r_x^\top X}\Big|Z\Big]\ \mathbb{E}\Big[e^{is_2 r_y^\top Y}\Big|Z\Big] \quad \text{a.s.}$$

Since the parameter set is countable, we may choose versions of the conditional characteristic functions such that the above identity holds simultaneously for all $(r_x, r_y, s_1, s_2) \in \mathcal{D}_x \times \mathcal{D}_y \times \mathbb{Q}^2$ and for μ_Z-a.e. z.

Fix such a z and define

$$\Phi_z(t_x, t_y) := \varphi_{X,Y|Z}(t_x, t_y \mid z) - \varphi_{X|Z}(t_x \mid z)\,\varphi_{Y|Z}(t_y \mid z), \qquad (t_x, t_y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}.$$

For every $(r_x, r_y, s_1, s_2) \in \mathcal{D}_x \times \mathcal{D}_y \times \mathbb{Q}^2$, we have

$$\Phi_z(s_1 r_x, s_2 r_y) = 0.$$

By continuity of characteristic functions in their frequency arguments, this extends from $\mathcal{D}_x \times \mathcal{D}_y \times \mathbb{Q}^2$ to all

$$(r_x, r_y, s_1, s_2) \in \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1} \times \mathbb{R}^2.$$

Now let $(t_x, t_y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ be arbitrary. If $t_x \neq 0$, write

$$t_x = \|t_x\| \frac{t_x}{\|t_x\|} =: s_1 r_x.$$

If $t_x = 0$, take $s_1 = 0$ and arbitrary $r_x \in \mathbb{S}^{d_x-1}$. Do the same for t_y. Thus there exist $r_x \in \mathbb{S}^{d_x-1}$, $r_y \in \mathbb{S}^{d_y-1}$, and $s_1, s_2 \in \mathbb{R}$ such that

$$t_x = s_1 r_x, \qquad t_y = s_2 r_y.$$

Hence

$$\Phi_z(t_x, t_y) = 0 \qquad \forall (t_x, t_y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}.$$

Since z was arbitrary on a set of full μ_Z-measure, this factorization holds for μ_Z-a.e. z:

$$\varphi_{X,Y|Z}(t_x, t_y \mid z) = \varphi_{X|Z}(t_x \mid z)\,\varphi_{Y|Z}(t_y \mid z) \qquad \forall (t_x, t_y) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}.$$

By the uniqueness theorem for characteristic functions, for μ_Z-a.e. z,

$$P_{X,Y|Z}(\cdot, \cdot \mid z) = P_{X|Z}(\cdot \mid z) \otimes P_{Y|Z}(\cdot \mid z),$$

which is precisely $X \perp\!\!\!\perp Y \mid Z$. This contradicts $X \not\perp\!\!\!\perp Y \mid Z$. Therefore, there must exist at least one projection pair $(r_x^\star, r_y^\star)$ such that the projected variables are conditionally dependent given Z. Under the projected CIT characterization, this is equivalent to $\rho(r_x^\star, r_y^\star) > 0$. □

E.2.2 Continuity of the projected population CIT index

Theorem E.5 (Continuity of the projected population CIT index)**.** *Under Assumption 2.3, the map*

$$(r_x, r_y) \longmapsto \rho(r_x, r_y)$$

is continuous on $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$.

We prove continuity of the projected population CIT index through the chain

$$(r_x, r_y) \xrightarrow{\text{(i)}} \big(G_X(Z, r_x, \cdot),\, G_Y(Z, r_y, \cdot)\big) \xrightarrow{\text{(ii)}} (U, V, W) \xrightarrow{\text{(iii)}} \rho.$$

Throughout this subsection, we write

$$\mathcal{Z}_+ := \{z \in \mathcal{S}_Z : f_Z(z) > 0\}.$$

Since $\mu_Z(\mathcal{Z}_+) = 1$, all almost-sure statements with respect to Z may be restricted to $\mathcal{Z}_+$ without loss. Continuity on $\mathcal{Z}_+$ is understood relative to the subspace topology.

Lemma E.6 (Jointly measurable versions of projected conditional CDFs)**.** *Under Assumption 2.3, define*

$$G_X(z, r, t) := \frac{\int_{\mathbb{R}^{d_x}} \mathbf{1}\{r^\top x \le t\}\, f_{X,Z}(x, z)\, dx}{f_Z(z)},$$

for $z \in \mathcal{Z}_+$, $r \in \mathbb{S}^{d_x-1}$, $t \in \mathbb{R}$, *and define* G_Y *analogously using* $f_{Y,Z}$. *Then:*

(a) G_X *and* G_Y *are jointly measurable on* $\mathcal{Z}_+ \times \mathbb{S}^{d_x-1} \times \mathbb{R}$ *(resp.* $\mathcal{Z}_+ \times \mathbb{S}^{d_y-1} \times \mathbb{R}$*).*

(b) *For* μ_Z*-a.e.* $z \in \mathcal{Z}_+$, $G_X(z, r, t)$ *is a regular version of* $\mathbb{P}(r^\top X \le t \mid Z = z)$, *and similarly for* G_Y.

Proof. We prove the claim for G_X; the proof for G_Y is identical.

For part (a), the map $(x, z, r, t) \mapsto \mathbf{1}\{r^\top x \le t\}\, f_{X,Z}(x, z)$ is Borel measurable. Integrating out x preserves measurability by the Fubini–Tonelli theorem. On $\mathcal{Z}_+$, $f_Z(z) > 0$ by definition, so the ratio G_X is well-defined and jointly measurable.

For part (b), Bayes' formula gives, for μ_Z-a.e. $z \in \mathcal{Z}_+$ and every Borel set $A \subseteq \mathbb{R}^{d_x}$,

$$\mathbb{P}(X \in A \mid Z = z) = \frac{\int_A f_{X,Z}(x, z)\, dx}{f_Z(z)}.$$

Taking $A = \{x \in \mathbb{R}^{d_x} : r^\top x \le t\}$ yields $G_X(z, r, t)$. □

Lemma E.7 (Joint continuity of projected conditional CDFs)**.** *Under Assumption 2.3, G_X is jointly continuous on $\mathcal{Z}_+ \times \mathbb{S}^{d_x-1} \times \mathbb{R}$, and G_Y is jointly continuous on $\mathcal{Z}_+ \times \mathbb{S}^{d_y-1} \times \mathbb{R}$, where continuity in the z coordinate is understood relative to $\mathcal{Z}_+$.*

Proof. We prove the claim for G_X; the proof for G_Y is identical.

Let $(z_m, r_m, t_m) \to (z, r, t)$ with $z_m, z \in \mathcal{Z}_+$. Define

$$N_m := \int_{\mathbb{R}^{d_x}} \mathbf{1}\{r_m^\top x \le t_m\}\, f_{X,Z}(x, z_m)\, dx.$$

For every fixed $x \notin \{x : r^\top x = t\}$ (a set of Lebesgue measure zero), when $(r_m, t_m) \to (r, t)$,

$$\mathbf{1}\{r_m^\top x \le t_m\} \to \mathbf{1}\{r^\top x \le t\}.$$

Since $f_{X,Z}(x, z_m) \to f_{X,Z}(x, z)$ for a.e. x by Condition G, the integrand converges pointwise for a.e. x.

Since $z_m \to z$ and $z_m, z \in \mathcal{Z}_+ \subseteq \mathcal{S}_Z$, the set $K := \{z\} \cup \{z_m : m \ge 1\}$ is a compact subset of $\mathcal{S}_Z$. Condition G thus provides an integrable dominating function $M_{X,K}(x)$ such that $f_{X,Z}(x, z_m) \le M_{X,K}(x)$ for every m and a.e. x. By the dominated convergence theorem,

$$N_m \to \int_{\mathbb{R}^{d_x}} \mathbf{1}\{r^\top x \le t\}\, f_{X,Z}(x, z)\, dx.$$

Since $z_m \to z \in \mathcal{Z}_+$, continuity of f_Z (Condition H) gives $f_Z(z_m) \to f_Z(z) > 0$. Therefore $G_X(z_m, r_m, t_m) \to G_X(z, r, t)$. □

Proposition E.8 (Almost sure convergence of Rosenblatt-transformed variables)**.** *Under Assumption 2.3, let $r_x^{(m)} \to r_x$ in $\mathbb{S}^{d_x-1}$ and $r_y^{(m)} \to r_y$ in $\mathbb{S}^{d_y-1}$. Define*

$$U_m := G_X\big(Z, r_x^{(m)}, (r_x^{(m)})^\top X\big), \qquad U := G_X\big(Z, r_x, r_x^\top X\big),$$

and analogously

$$V_m := G_Y\big(Z, r_y^{(m)}, (r_y^{(m)})^\top Y\big), \qquad V := G_Y\big(Z, r_y, r_y^\top Y\big).$$

Let $W := F_Z(Z)$. Then $(U_m, V_m, W) \to (U, V, W)$ a.s.

Proof. Fix ω such that $Z(\omega) \in \mathcal{Z}_+$; this holds $\mathbb{P}$-a.s. since $\mu_Z(\mathcal{Z}_+) = 1$. Set $z := Z(\omega)$, $x := X(\omega)$, $y := Y(\omega)$. Let $t_m^X := (r_x^{(m)})^\top x$ and $t^X := r_x^\top x$; similarly $t_m^Y := (r_y^{(m)})^\top y$ and $t^Y := r_y^\top y$. By continuity of the inner product, $t_m^X \to t^X$ and $t_m^Y \to t^Y$.

Since $z \in \mathcal{Z}_+$, G_X is defined at $(z, r_x^{(m)}, t_m^X)$ for every m. By Lemma E.7,

$$U_m(\omega) = G_X(z, r_x^{(m)}, t_m^X) \to G_X(z, r_x, t^X) = U(\omega),$$

and similarly $V_m(\omega) \to V(\omega)$. Since W does not depend on the projection directions, the result follows. □

Proof of Theorem E.5. Let $(r_x^{(m)}, r_y^{(m)}) \to (r_x, r_y)$ in $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$. For each m, let $(U_m^{(i)}, V_m^{(i)}, W^{(i)})$, $i = 1, 2$, denote two independent copies of the Rosenblatt-transformed triple corresponding to directions $(r_x^{(m)}, r_y^{(m)})$, and let $(U^{(i)}, V^{(i)}, W^{(i)})$, $i = 1, 2$, be the corresponding copies for (r_x, r_y).

By the population form of the CIT index,

$$\rho(r_x, r_y) = c_0 \, \mathbb{E}\Big[h\big((U^{(1)}, V^{(1)}, W^{(1)}), (U^{(2)}, V^{(2)}, W^{(2)})\big)\Big],$$

where

$$h\big((u_1, v_1, w_1), (u_2, v_2, w_2)\big) := \kappa(u_1, u_2)\, \kappa(v_1, v_2)\, e^{-|w_1 - w_2|},$$

with

$$\kappa(a, b) := e^{-|a-b|} + e^{-a} + e^{a-1} + e^{-b} + e^{b-1} + 2e^{-1} - 4.$$

The kernel h is bounded and continuous on $[0, 1]^3 \times [0, 1]^3$.

By Proposition E.8, for each copy,

$$(U_m^{(i)}, V_m^{(i)}, W^{(i)}) \to (U^{(i)}, V^{(i)}, W^{(i)}) \quad \text{a.s.}$$

Hence

$$h\big((U_m^{(1)}, V_m^{(1)}, W^{(1)}), (U_m^{(2)}, V_m^{(2)}, W^{(2)})\big) \to h\big((U^{(1)}, V^{(1)}, W^{(1)}), (U^{(2)}, V^{(2)}, W^{(2)})\big) \quad \text{a.s.}$$

Since h is bounded, the bounded convergence theorem yields

$$\rho(r_x^{(m)}, r_y^{(m)}) \to \rho(r_x, r_y).$$

Therefore ρ is continuous. □

E.2.3 Positive probability of capturing signal

Corollary E.9 (Positive projection-capture probability)**.** *Suppose $X \not\perp\!\!\!\perp Y \mid Z$ and Assumption 2.3 holds. Then there exist a measurable set $N \subseteq \mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$ and constants $\epsilon_1, \pi > 0$ such that*

$$\rho(r_x, r_y) \geq \epsilon_1 \qquad \textit{for all } (r_x, r_y) \in N,$$

and

$$\mathbb{P}\big((R_x, R_y) \in N\big) = \pi > 0,$$

where $R_x \sim \mathrm{Unif}(\mathbb{S}^{d_x-1})$ and $R_y \sim \mathrm{Unif}(\mathbb{S}^{d_y-1})$ are independent.

Proof. By Lemma E.4, there exists $(r_x^\star, r_y^\star)$ such that $\rho(r_x^\star, r_y^\star) > 0$. By Theorem E.5, there exist an open neighborhood $N \ni (r_x^\star, r_y^\star)$ in $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$ and a constant $\epsilon_1 > 0$ such that

$$\rho(r_x, r_y) \geq \epsilon_1 \qquad \text{for all } (r_x, r_y) \in N.$$

Since the product of the uniform surface measures on $\mathbb{S}^{d_x-1} \times \mathbb{S}^{d_y-1}$ assigns strictly positive mass to every nonempty open set, letting

$$\pi := \mathbb{P}\big((R_x, R_y) \in N\big)$$

gives $\pi > 0$. □

E.2.4 Projection capture and reference tail control

Lemma E.10 (Reference critical-value bound)**.** *Let h be the kernel defined in the proof of Theorem E.5, and set*

$$L := c_0 \max_{s \in [0,1]^3} h(s, s) \in (0, \infty).$$

For integers $n, B \geq 1$ and $u \in (0, 1)$ satisfying $\lfloor (B+1)u \rfloor \geq 1$, the Monte Carlo reference threshold defined in Appendix C.1 satisfies, for every $x \geq 0$,

$$\mathbb{P}\big\{q_{n,u} \geq L\big(2 + 4\log(2/u) + 4x\big)\big\} \leq e^{-x}.$$

Proof. The kernel κ in the proof of Theorem E.5 is the uniform-centering of the Laplace kernel:

$$\kappa(a,b) = e^{-|a-b|} - \int_0^1 e^{-|a-t|}\,dt - \int_0^1 e^{-|b-t|}\,dt + \int_0^1\int_0^1 e^{-|s-t|}\,ds\,dt.$$

Thus κ is positive semidefinite and $\int_0^1 \kappa(a,b)\,db = 0$. Since $e^{-|w-w'|}$ is also positive semidefinite and $c_0 > 0$, the product kernel $c_0 h$ is positive semidefinite, bounded, and centered under $\mathrm{Unif}(0,1)^3$:

$$\int_{[0,1]^3} c_0 h(s,t)\,dt = 0, \qquad 0 \le c_0 h(s,s) \le L.$$

Let Ψ be a Hilbert-space feature map of $c_0 h$. For i.i.d. reference triples $S_i^* = (U_i^*, V_i^*, W_i^*) \sim \mathrm{Unif}(0,1)^3$, centering yields $\mathbb{E}\Psi(S_i^*) = 0$, while $\|\Psi(S_i^*)\|^2 \le L$. By Algorithm 2,

$$n\widehat{\rho}_n^* = \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \Psi(S_i^*) \right\|^2.$$

Put $R_n = \sqrt{n\widehat{\rho}_n^*}$. Independence and centering give $\mathbb{E}R_n \le \sqrt{\mathbb{E}R_n^2} \le \sqrt{L}$. Replacing one reference triple changes R_n by at most $2\sqrt{L/n}$. McDiarmid's inequality therefore yields

$$\mathbb{P}\Big\{ R_n \ge \sqrt{L} + \sqrt{2Lx} \Big\} \le e^{-x}, \qquad x \ge 0.$$

As $(\sqrt{L} + \sqrt{2Lx})^2 \le L(2+4x)$,

$$\mathbb{P}\{ n\widehat{\rho}_n^* \ge L(2+4x) \} \le e^{-x}.$$

Let $m = \lfloor (B+1)u \rfloor \ge 1$. If $q_{n,u} \ge z$, at least m of the B reference statistics are at least z. Hence, by Markov's inequality and $m \ge (B+1)u/2$,

$$\begin{aligned}\mathbb{P}(q_{n,u} \ge z) &\le \frac{1}{m} \sum_{b=1}^B \mathbb{P}(n\widehat{\rho}_n^{*(b)} \ge z) \\ &\le \frac{2}{u}\,\mathbb{P}(n\widehat{\rho}_n^* \ge z).\end{aligned}$$

Taking $z = L\{2 + 4\log(2/u) + 4x\}$ completes the proof. □

Fix $\alpha \in (0,1)$ and a distribution $P \in H_1$ satisfying Assumption 2.3. Let N, ϵ_1, π be as in Corollary E.9. Define the average estimation error on N by

$$\bar{\delta}_n := \frac{1}{\pi} \int_N \mathbb{P}_P\{ |\widehat{\rho}_n(\omega) - \rho(\omega)| > \epsilon_1/2 \}\; d\mathbb{P}_{\mathcal{O}}(\omega).$$

For integers $k, B \ge 1$ satisfying $1/(B+1) \le \alpha/k$, let $\Omega_1, \ldots, \Omega_k$ be i.i.d. projection pairs with law $\mathbb{P}_{\mathcal{O}}$, independent of the observed data and the shared Monte Carlo reference pool. Set $G_\ell = \{\Omega_\ell \in N\}$. On $\bigcup_{\ell=1}^k G_\ell$, let J denote the first index for which G_J occurs. Then

$$\begin{aligned}&\mathbb{P}_P\left\{ \bigcup_{\ell=1}^k G_\ell,\ |\widehat{\rho}_n(\Omega_J) - \rho(\Omega_J)| > \epsilon_1/2 \right\} \\ &= \sum_{j=1}^k (1-\pi)^{j-1} \int_N \mathbb{P}_P\{ |\widehat{\rho}_n(\omega) - \rho(\omega)| > \epsilon_1/2 \}\; d\mathbb{P}_{\mathcal{O}}(\omega) \\ &= \{1 - (1-\pi)^k\}\bar{\delta}_n \le \bar{\delta}_n.\end{aligned}$$

On $\bigcup_{\ell=1}^k G_\ell$, if $|\widehat{\rho}_n(\Omega_J) - \rho(\Omega_J)| \le \epsilon_1/2$, then $n\widehat{\rho}_n(\Omega_J) \ge n\epsilon_1/2$. If also $q_{n,\alpha/k} < n\epsilon_1/2$, the plus-one rule in Appendix C.1 rejects at projection J. Consequently, the Type II error satisfies

$$\beta_n \le (1-\pi)^k + \bar{\delta}_n + \mathbb{P}(q_{n,\alpha/k} \ge n\epsilon_1/2).$$

For $n\epsilon_1/2 \geq 2L$, the argument in Lemma E.10, with $u = \alpha/k$, gives

$$\mathbb{P}(q_{n,\alpha/k} \geq n\epsilon_1/2) \leq \frac{2e^{1/2}}{\alpha}\, k \exp\Big(-\frac{n\epsilon_1}{8L}\Big).$$

Thus, whenever the reference-resolution condition holds and $n\epsilon_1/2 \geq 2L$,

$$\beta_n \leq (1-\pi)^k + \bar{\delta}_n + \frac{2e^{1/2}}{\alpha}\, k \exp\Big(-\frac{n\epsilon_1}{8L}\Big). \tag{4}$$

If $\widehat{\rho}_n(\omega) \xrightarrow{P} \rho(\omega)$ for $\mathbb{P}_{\mathcal{O}}$-almost every $\omega \in N$, then bounded convergence implies $\bar{\delta}_n \to 0$.

E.2.5 Power for fixed and growing projection budgets

Proof of Theorem 2.5. The assumed scalar CIT fixed-alternative consistency conditions imply $\widehat{\rho}_n(\omega) \xrightarrow{P} \rho(\omega)$ for $\mathbb{P}_{\mathcal{O}}$-almost every $\omega \in N$. Thus $\bar{\delta}_n \to 0$. For fixed k and deterministic B_n satisfying $1/(B_n+1) \leq \alpha/k$ eventually, applying (4) with $B = B_n$ gives

$$\limsup_{n\to\infty} \beta_n \leq (1-\pi)^k.$$

□

Corollary E.11 (Power with a growing number of projections)**.** *Fix $\alpha \in (0,1)$ and a distribution $P \in H_1$ satisfying Assumption 2.3. Let N, ϵ_1, π be as in Corollary E.9. Assume that $\widehat{\rho}_n(\omega) \xrightarrow{P} \rho(\omega)$ for $\mathbb{P}_{\mathcal{O}}$-almost every $\omega \in N$. Let $(k_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ be deterministic sequences of positive integers satisfying $\log k_n = o(n)$ and*

$$\frac{1}{B_n+1} \leq \frac{\alpha}{k_n} \qquad \textit{for all sufficiently large } n.$$

Then the Type II error of Bonferroni-max aggregation satisfies

$$\beta_n \leq (1-\pi)^{k_n} + o(1).$$

In particular, if $k_n \to \infty$, then $\beta_n \to 0$.

Proof. The consistency assumption gives $\bar{\delta}_n \to 0$. Since $\log k_n = o(n)$ and $\epsilon_1, L > 0$ are fixed,

$$k_n \exp\Big(-\frac{n\epsilon_1}{8L}\Big) \longrightarrow 0.$$

Applying (4) with $k = k_n$ and $B = B_n$ completes the proof. □

The power bounds above are pointwise in the alternative distribution. For fixed k, the same capture argument yields a uniform projection-capture bound over an alternative class if capture probabilities have a common positive lower bound and projected detection-failure probabilities at level α/k vanish uniformly over the class and the corresponding informative regions.

E.3 Formal results and proofs for top-r aggregation

Throughout, g_r, L_r, $t_{\alpha,k,r}$, and τ_λ are as in Section 3.1.

E.3.1 Conditional binomial representation

Let $D_n := \{(X_i, Y_i, Z_i)\}_{i=1}^n$ denote the observed data, $\mathcal{R}_n$ the shared Monte Carlo reference pool, and $\mathcal{F}_n := \sigma(D_n, \mathcal{R}_n)$. For $u \in (0,1)$, define

$$N_n(u) := \sum_{\ell=1}^{k} \mathbf{1}\{p_n^{(\ell)} \leq u\}, \qquad \phi_n^{(r)}(u) = \mathbf{1}\{N_n(u) \geq r\}, \qquad Q_n(u) := \mathbb{P}_P\{p_n(\Omega) \leq u \mid \mathcal{F}_n\},$$

where Ω is a fresh projection pair independent of $\mathcal{F}_n$.

Lemma E.12 (Conditional binomial representation)**.** *For every $u \in (0,1)$ and $r \in \{1,\dots,k\}$,*

$$N_n(u) \mid \mathcal{F}_n \sim \mathrm{Bin}\{k, Q_n(u)\}, \qquad \mathbb{P}_P\{\phi_n^{(r)}(u) = 1\} = \mathbb{E}_P\big[g_r\{Q_n(u)\}\big],$$

and $\mathbb{E}_P Q_n(u) = \mathbb{P}_P\{p_n(\Omega) \le u\}$.

Proof. Each projected p-value is the same measurable function of $(D_n, \mathcal{R}_n)$ and its own projection pair. The projection pairs are i.i.d. and independent of $\mathcal{F}_n$, so the indicators $\mathbf{1}\{p_n^{(\ell)} \le u\}$ are conditionally i.i.d. $\mathrm{Bernoulli}\{Q_n(u)\}$ given $\mathcal{F}_n$. Taking expectations gives the other two identities. □

Since $g_r'(q) = k\,\mathbb{P}\{\mathrm{Bin}(k-1,q) = r-1\}$ increases on $[0,(r-1)/(k-1)]$, g_r is convex there. For $k = 1000$, $\alpha = 0.05$, $\lambda \ge 1$, and $r \in \{2,5,10\}$, the threshold τ_λ lies in this interval. When Q_n stays in this range, Jensen's inequality gives $\mathbb{E}_P g_r(Q_n) \ge g_r(\mathbb{E}_P Q_n)$: fluctuations of Q_n raise the rejection probability, which is why the Beta threshold requires a correction.

E.3.2 Type I control

Theorem E.13 (Markov top-r Type I control)**.** *Fix $P \in H_0$, $\alpha \in (0,1)$, $k \ge 1$, and $r \in \{1,\dots,k\}$. If a fresh projected p-value is marginally asymptotically valid, that is, $\limsup_n \mathbb{P}_P\{p_n(\Omega) \le u\} \le u$ for every fixed $u \in (0,1)$, then*

$$\limsup_{n\to\infty} \mathbb{P}_P\{\phi_n^{(r)}(\alpha r/k) = 1\} \le \alpha.$$

Under the conditions of Theorem 2.2, marginal validity follows from the single-projection bound in Appendix E.1.

Proof. With $u = \alpha r/k$, Markov's inequality and Lemma E.12 give

$$\mathbb{P}_P\{N_n(u) \ge r\} \le \frac{\mathbb{E}_P N_n(u)}{r} = \frac{k}{r}\,\mathbb{P}_P\{p_n(\Omega) \le u\},$$

whose limit superior is at most $ku/r = \alpha$. □

Proof of Theorem 3.2. Write $\tau = \tau_\lambda$, $Q_n = Q_n(\tau)$, $\mu_n = \mathbb{E}_P Q_n$, and $v_n = \mathrm{Var}_P(Q_n)$. Since $g_r'' \le L_r$ on $[0,1]$, Lemma E.12 and a second-order Taylor expansion of g_r around μ_n give

$$\mathbb{P}_P\{\phi_n^{(r)}(\tau) = 1\} = \mathbb{E}_P[g_r(Q_n)] \le g_r(\mu_n) + \frac{L_r}{2} v_n.$$

Marginal validity gives $\limsup_n \mu_n \le \tau$. Since g_r is continuous and increasing, Assumption 3.1 yields

$$\limsup_{n\to\infty} \mathbb{P}_P\{\phi_n^{(r)}(\tau) = 1\} \le g_r(\tau) + \frac{L_r}{2} \limsup_{n\to\infty} v_n \le \alpha.$$

□

E.3.3 Gaussian-residual verification

Consider i.i.d. observations with

$$Z \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}], \qquad X = aZ + \varepsilon_X, \qquad Y = bZ + \varepsilon_Y,$$

where $\varepsilon_X \sim N(0, I_{d_x})$ and $\varepsilon_Y \sim N(0, I_{d_y})$ are independent of each other and of Z, and $a \in \mathbb{R}^{d_x}$, $b \in \mathbb{R}^{d_y}$ are fixed. CIT uses Algorithm 2 with a Gaussian kernel and bandwidth $h_n = c\,\widehat{\sigma}_Z n^{-1/5}$, where $c > 0$ is fixed and $\widehat{\sigma}_Z$ is the sample standard deviation of Z. We analyze this rule without the implementation's numerical guard that resets h_n to 1 when $h_n < 10^{-8}$, which is inactive at the sample sizes considered. Projection pairs are i.i.d. uniform on $\mathcal{O}$ and independent of the data and of a shared reference pool of fixed size B, generated afresh at each n. Expectations and variances include the reference-pool randomness.

Proposition E.14 (Gaussian-residual verification)**.** *In this model, fix $k \geq 2$, $r \in \{2, \ldots, k\}$, $\alpha \in (0,1)$, and $\lambda \geq 1$ in advance, and let $\tau = \tau_\lambda$, $D = d_x d_y$,*

$$\mu_B = \frac{\lfloor (B+1)\tau \rfloor}{B+1}, \qquad s_B = \frac{\mu_B(1-\mu_B)}{B+2}, \qquad V_B(D) = s_B + \frac{\mu_B(1-\mu_B) - s_B}{D}.$$

Then $\mathbb{E}_P Q_n(\tau) \to \mu_B \leq \tau$ and $\limsup_{n\to\infty} \mathrm{Var}_P\{Q_n(\tau)\} \leq V_B(D)$. Consequently, if $V_B(D) \leq 2\{\alpha - g_r(\tau)\}/L_r$, then Assumption 3.1 holds and $\limsup_{n\to\infty} \mathbb{P}_P\{\phi_n^{(r)}(\tau) = 1\} \leq \alpha$.

The term s_B reflects reference-pool variability. The second term reflects dependence between projections that share the data; it decays as $1/D$ because independent random directions are nearly orthogonal in high dimension. With $a = b = \mathbf{1}_d$, the model is the N1 design of Appendix D.1 with bounded Z. For $k = 1000$, $\alpha = 0.05$, $\lambda = 1.5$, and $B = 500{,}000$, the sufficient condition holds for $r \in \{5, 10\}$ at $d \in \{50, 200\}$ and for $r = 2$ at $d = 200$, but not at $d = 10$. Since $V_B(D)$ is an upper bound, its failure does not imply that the level is exceeded; Appendix D.4 reports empirical diagnostics.

Proof of Proposition E.14. Let $I_{n,j} = \mathbf{1}\{p_n(\Omega_j) \leq \tau\}$ for two independent fresh projection pairs Ω_1, Ω_2. By Lemma E.12, the indicators are conditionally independent given $\mathcal{F}_n$ with mean $Q_n(\tau)$, so $\mathrm{Var}_P\{Q_n(\tau)\} = \mathrm{Cov}_P(I_{n,1}, I_{n,2})$. Set $m_B = \lfloor (B+1)\tau \rfloor$. If $m_B = 0$, both indicators vanish and the claims are immediate. Otherwise, the plus-one rule of Appendix C.1 rejects when $n\widehat{\rho}_n(\Omega_j) > q_{n,\tau}$, the $(B - m_B + 1)$st ascending reference statistic.

Fix directions $\omega_j = (u_j, v_j)$ and write $\rho_x = u_1^\top u_2$ and $\rho_y = v_1^\top v_2$. By Lemmas E.16 and E.17 and independence of the reference pool, the two observed statistics and the B reference statistics converge jointly to limits with a common continuous CDF F_∞. Let C_R be the limiting reference order statistic and $q_R = 1 - F_\infty(C_R)$. Then

$$q_R \sim \mathrm{Beta}(m_B, B + 1 - m_B), \qquad \mathbb{E}q_R = \mu_B, \qquad \mathrm{Var}(q_R) = s_B.$$

Conditionally on C_R, the limiting indicators have mean q_R and, by Lemma E.17, covariance at most $\rho_x^2 \rho_y^2 q_R(1 - q_R)$. The law of total covariance bounds their covariance by $s_B + \rho_x^2 \rho_y^2 \, \mathbb{E}\{q_R(1 - q_R)\}$. Since $\mathbb{E}(\rho_x^2 \rho_y^2) = D^{-1}$ for independent uniform directions and $\mathbb{E}\{q_R(1-q_R)\} = \mu_B(1-\mu_B) - s_B$, averaging over directions gives the stated limits. Bounded convergence applies because the indicator moments converge for every fixed pair of directions and the limiting cutoff has no ties. The final claim follows from Theorem 3.2. □

E.3.4 Power

Theorem E.15 (Power of top-r aggregation)**.** *Fix $P \in H_1$ satisfying Assumption 2.3, and let N, ϵ_1, and π be as in Corollary E.9. Fix $k \geq 1$, $r \in \{1, \ldots, k\}$, and $\tau \in (0,1)$, and let $(B_n)_{n\geq 1}$ be deterministic with $1/(B_n + 1) \leq \tau$ for all sufficiently large n. Assume that the projection pairs are independent and that, for $\mathbb{P}_\mathcal{O}$-almost every $\omega \in N$, the projected CIT satisfies the fixed-alternative consistency conditions of Cai et al. [4, Theorem 5]. Then the Type II error of $\phi_n^{(r)}(\tau)$ satisfies*

$$\limsup_{n\to\infty} \beta_n^{(r)}(\tau) \leq \mathbb{P}\{\mathrm{Bin}(k, \pi) \leq r - 1\},$$

which equals $(1-\pi)^k$ for $r = 1$.

Proof. Let $M_k = \sum_{\ell=1}^k \mathbf{1}\{\Omega_\ell \in N\} \sim \mathrm{Bin}(k, \pi)$. On $\{M_k \geq r\}$, let $J_1, \ldots, J_r$ index the first r projections in N. If $|\widehat{\rho}_n(\Omega_{J_j}) - \rho(\Omega_{J_j})| \leq \epsilon_1/2$ for $j \leq r$ and $q_{n,\tau} < n\epsilon_1/2$, all r selected projections reject. With the average estimation error $\bar{\delta}_n$ of Appendix E.2, the argument leading to (4) gives

$$\beta_n^{(r)}(\tau) \leq \mathbb{P}\{\mathrm{Bin}(k, \pi) \leq r - 1\} + r\bar{\delta}_n + \mathbb{P}(q_{n,\tau} \geq n\epsilon_1/2).$$

Projected CIT consistency gives $\bar{\delta}_n \to 0$, and the resolution condition and Lemma E.10 give $\mathbb{P}(q_{n,\tau} \geq n\epsilon_1/2) \to 0$. □

E.3.5 Technical lemmas for Gaussian-residual verification

Throughout, we use the model and bandwidth of Appendix E.3.3. Lemma E.16 shows that the estimated-CDF statistic is asymptotically equivalent to its oracle version, via a kernel representation, uniform bounds on the conditional-CDF errors, and a tensor expansion of the error. Lemma E.17 bounds the dependence between the rejection events of two projections.

Lemma E.16 (Estimated-CDF oracle approximation)**.** *For every fixed projection pair ω, let $\widehat{T}_n(\omega) = n\widehat{\rho}_n(\omega)$ denote the estimated-CDF CIT statistic and $T_n^\circ(\omega)$ the same statistic computed from the exact Rosenblatt transforms. Under this setup,*

$$\widehat{T}_n(\omega) - T_n^\circ(\omega) \xrightarrow{P} 0.$$

Proof. Write $A = \sqrt{3}$ and $h_0 = n^{-1/5}$. For a fixed $\omega = (u, v)$, the oracle transforms are

$$U_i = \Phi(u^\top \varepsilon_{X,i}), \qquad V_i = \Phi(v^\top \varepsilon_{Y,i}), \qquad W_i = F_Z(Z_i).$$

Conditional on $\mathcal{Z}_n = (Z_1, \dots, Z_n)$, the U and V arrays are independent arrays of i.i.d. uniform variables. Let

$$w_{ij} = \frac{\exp\{-(Z_j - Z_i)^2/(2h_n^2)\}}{\sum_{l=1}^n \exp\{-(Z_l - Z_i)^2/(2h_n^2)\}}, \qquad \widehat{U}_i = \sum_j w_{ij} \mathbf{1}\{u^\top X_j \le u^\top X_i\},$$

with $\widehat{V}_i$ defined analogously and $\widehat{W}_i = n^{-1} \sum_j \mathbf{1}\{Z_j \le Z_i\}$. The term $j = i$ is retained.

Kernel representation. Define $J_x(s) = \mathbf{1}\{x \le s\} - s$ on $[0, 1]$ and the operator $(\mathcal{K}f)(s) = \int_0^1 e^{-|s-t|} f(t)\, dt$ on $H = L^2[0, 1]$. For the uniformly centered CIT kernel κ, integration by parts in both arguments gives

$$\kappa(x, y) = \langle J_x, (2I - \mathcal{K}) J_y \rangle_H, \qquad \|J_x - J_y\|_H^2 = |x - y|, \qquad \mathbb{E} J_U = 0.$$

The operator $A_0 = 2I - \mathcal{K}$ is bounded and positive. Let ψ be a feature map for $\ell(w, w') = e^{-|w-w'|}$. Then $\|\psi(w)\| = 1$ and $\|\psi(w) - \psi(w')\|^2 \le 2|w - w'|$. In $H \otimes H \otimes H_W$, put

$$S_n = \frac{1}{\sqrt{n}} \sum_i J_{U_i} \otimes J_{V_i} \otimes \psi(W_i),$$
$$\widehat{S}_n = \frac{1}{\sqrt{n}} \sum_i J_{\widehat{U}_i} \otimes J_{\widehat{V}_i} \otimes \psi(\widehat{W}_i).$$

The CIT kernel representation yields

$$T_n^\circ = c_0 \langle S_n, (A_0 \otimes A_0 \otimes I) S_n \rangle,$$
$$\widehat{T}_n = c_0 \langle \widehat{S}_n, (A_0 \otimes A_0 \otimes I) \widehat{S}_n \rangle.$$

The bounded centered oracle summands imply $\|S_n\| = O_P(1)$. It remains to show $\|\widehat{S}_n - S_n\| = o_P(1)$.

Conditional CDF error. Let $\Delta_{Z,n} = \sup_z |F_{n,Z}(z) - F_Z(z)|$ and $\delta_i = \min\{Z_i + A, A - Z_i\}$. By the Dvoretzky–Kiefer–Wolfowitz inequality and $\widehat{\sigma}_Z \to 1$, there are constants $c_-, c_+, C > 0$, depending on the fixed c, and design events $\mathcal{E}_n$ with $\mathbb{P}(\mathcal{E}_n) \to 1$ such that on $\mathcal{E}_n$,

$$\Delta_{Z,n} \le C\sqrt{\frac{\log n}{n}}, \qquad c_- h_0 \le h_n \le c_+ h_0.$$

Conditioning on such a design makes all w_{ij} deterministic. For bounded-variation f, $|\int f\, d(F_{n,Z} - F_Z)| \le \Delta_{Z,n}\, \mathrm{TV}(f)$. Applying this bound to Gaussian-kernel moments gives, uniformly in i,

$$\max_j w_{ij} \le \frac{C}{nh_0}, \qquad \sum_j w_{ij}^2 \le \frac{C}{nh_0},$$

$$\left| \sum_j w_{ij}(Z_j - Z_i) \right| \le C\{h_0 e^{-\delta_i^2/(Ch_0^2)} + \Delta_{Z,n}\}, \quad \sum_j w_{ij}(Z_j - Z_i)^2 \le Ch_0^2.$$

Indeed, for a constant $c_{\mathrm{den}} > 0$ depending on the fixed bandwidth coefficient, the normalized empirical kernel denominator satisfies, uniformly in i on $\mathcal{E}_n$,

$$\frac{1}{n}\sum_{j=1}^{n}\exp\left\{-\frac{(Z_j - Z_i)^2}{2h_n^2}\right\} \geq c_{\mathrm{den}} h_0,$$

including at the endpoints. For example, the population unnormalized first moment at $z \in [-A, A]$ equals

$$\frac{h_n^2}{2A}\left\{e^{-(z+A)^2/(2h_n^2)} - e^{-(A-z)^2/(2h_n^2)}\right\}.$$

Its empirical discrepancy is $O(h_n \Delta_{Z,n})$. The population second moment is $O(h_n^3)$ with empirical discrepancy $O(h_n^2 \Delta_{Z,n})$.

For the projected conditional CDF $F_u(x \mid z) = \Phi(x - (u^\top a)z)$, the first two derivatives in z are uniformly bounded. Taylor's formula and the preceding kernel moments give

$$\sup_x \left|\sum_j w_{ij} F_u(x \mid Z_j) - F_u(x \mid Z_i)\right| \leq C\{h_0^2 + h_0 e^{-\delta_i^2/(Ch_0^2)} + \Delta_{Z,n}\}.$$

The $O(h_0)$ endpoint bias is retained in this bound. At a deterministic threshold, weighted Hoeffding bounds the centered empirical sum by $2\exp\{-2e^2/\sum_j w_{ij}^2\}$. Applying this to a grid of n^3 quantiles of the conditional mixture CDF, then using monotonicity and a union bound over i, shows that the weighted empirical CDF error is at most $C\sqrt{\log n/(nh_0)} + n^{-3}$ uniformly over its evaluation argument, except with conditional probability at most $\eta_n = n^{-6}$. This includes the self-observation because the concentration bound is uniform in the argument before evaluation at $u^\top X_i$. The same statements hold for the V side.

Thus there are $\mathcal{Z}_n$-measurable envelopes $a_i \in [0, 1]$ such that, on $\mathcal{E}_n$,

$$\mathbb{P}\{\exists i : |\widehat{U}_i - U_i| > a_i \mid \mathcal{Z}_n\} \leq \eta_n, \qquad \mathbb{P}\{\exists i : |\widehat{V}_i - V_i| > a_i \mid \mathcal{Z}_n\} \leq \eta_n,$$
$$b_n := \frac{1}{n}\sum_i a_i^2 \leq C\left(h_0^3 + \frac{\log n}{nh_0}\right).$$

For the last bound, square the bias and concentration envelopes and use that the empirical mean of $e^{-\delta_i^2/(Ch_0^2)}$ is $O(h_0 + \Delta_{Z,n})$. In particular, $nb_n^2 \to 0$ on $\mathcal{E}_n$.

Tensor-error cancellation. Put $\Delta_i^U = J_{\widehat{U}_i} - J_{U_i}$ and $\Delta_i^V = J_{\widehat{V}_i} - J_{V_i}$. Expanding $\widehat{S}_n - S_n$ produces the four normalized sums with summands

$$\Delta_i^U \otimes J_{V_i} \otimes \psi(\widehat{W}_i), \qquad J_{U_i} \otimes \Delta_i^V \otimes \psi(\widehat{W}_i),$$
$$\Delta_i^U \otimes \Delta_i^V \otimes \psi(\widehat{W}_i), \qquad J_{U_i} \otimes J_{V_i} \otimes \{\psi(\widehat{W}_i) - \psi(W_i)\}.$$

For the first two sums, every off-diagonal inner-product expectation vanishes conditionally on $\mathcal{Z}_n$: the two entire residual arrays are independent and the untouched oracle feature is centered across observations. Their diagonal terms are bounded by $C(b_n^{1/2} + \eta_n)$ because $\|\Delta_i^U\|^2 = |\widehat{U}_i - U_i|$. For the mixed sum, $\Delta_i^U(s)$ is supported on $\{|U_i - s| \leq a_i\}$ on the good event. Independence of the oracle uniforms U_i, U_j for $i \neq j$ gives

$$\mathbb{E}\{|\langle \Delta_i^U, \Delta_j^U \rangle| \mid \mathcal{Z}_n\} \leq 4a_i a_j + \eta_n,$$

and likewise for V. Independence of the full U- and V-side arrays then bounds the mixed sum's conditional second moment by $C(b_n + nb_n^2 + n\eta_n^2)$. Finally, the ECDF feature difference has second moment at most $2\Delta_{Z,n}$, with vanishing off-diagonal terms by oracle centering. Consequently,

$$\mathbb{E}\{\|\widehat{S}_n - S_n\|^2 \mid \mathcal{Z}_n\} \leq C\{b_n^{1/2} + b_n + nb_n^2 + \Delta_{Z,n} + \eta_n + n\eta_n^2\} = o(1)$$

on $\mathcal{E}_n$. Conditional Markov's inequality proves $\|\widehat{S}_n - S_n\| = o_P(1)$ and therefore $\widehat{T}_n - T_n^\circ = o_P(1)$. □

Lemma E.17 (Two-projection Gaussian contraction)**.** *For fixed projection pairs $\omega_j = (u_j, v_j)$, $j = 1, 2$, the estimated CIT statistics converge jointly to $(\|G_1\|^2, \|G_2\|^2)$ for jointly Gaussian Hilbert-space elements with common squared-norm CDF F_∞. If $\rho_x = u_1^\top u_2$ and $\rho_y = v_1^\top v_2$, then, for any cutoff z, writing $q(z) = 1 - F_\infty(z)$,*

$$\left|\mathrm{Cov}\big(\mathbf{1}\{\|G_1\|^2 > z\}, \mathbf{1}\{\|G_2\|^2 > z\}\big)\right| \le \rho_x^2 \rho_y^2 q(z)\{1 - q(z)\}.$$

The distribution F_∞ is continuous and strictly increasing on $(0, \infty)$.

Proof. Let $\varphi(x) = A_0^{1/2} J_x$ and use the W-feature map ψ from the preceding proof. The oracle CIT statistic is the squared norm of the normalized sum of i.i.d. centered, bounded Hilbert elements

$$\xi_i^\omega = \sqrt{c_0}\, \varphi(U_i^\omega) \otimes \varphi(V_i^\omega) \otimes \psi(W_i).$$

The joint Hilbert-space central limit theorem for two fixed directions gives Gaussian elements G_1, G_2. The marginal covariance is nonzero, positive and trace class, so the squared norm has the law of a nondegenerate sum of weighted independent χ_1^2 variables. This law is continuous and strictly increasing on $(0, \infty)$. Joint convergence of the estimated statistics follows from Lemma E.16 and Slutsky's theorem.

Set $A_j = u_j^\top \varepsilon_X$ and $B_j = v_j^\top \varepsilon_Y$. The pairs (A_1, A_2) and (B_1, B_2) are independent Gaussian pairs with correlations ρ_x and ρ_y, and both are independent of W. For a scalar linear functional $H_j = \langle h_j, \xi^{\omega_j} \rangle$, separate centering gives $\mathbb{E}(H_j \mid A_j, W) = \mathbb{E}(H_j \mid B_j, W) = 0$. Writing h_p for the normalized Hermite polynomial of order p, the conditional Hermite expansion is

$$H_j = \sum_{p,q \ge 1} c_{pq,j}(W) h_p(A_j) h_q(B_j) \qquad \text{in } L^2.$$

Thus the common W remains in the coefficients, while separate centering excludes terms with $p = 0$ or $q = 0$. Mehler's identity and Cauchy–Schwarz yield

$$|\mathbb{E} H_1 H_2| \le |\rho_x \rho_y| (\mathbb{E} H_1^2\, \mathbb{E} H_2^2)^{1/2}.$$

The covariance of G_1, G_2 is that of the oracle summands, so the maximal correlation of their closed Gaussian linear spans is at most $|\rho_x \rho_y|$. The centered indicator $\mathbf{1}\{\|G_j\|^2 > z\} - q(z)$ is even in G_j and has Hermite rank at least two unless identically zero. Gebelein's inequality for Gaussian functionals [28, Theorem 2.3] therefore bounds the indicator covariance by the square of the maximal linear correlation times the marginal indicator variance $q(z)\{1 - q(z)\}$. □

E.4 Formal results and proofs for RP-GCM

E.4.1 Projected GCM conditions

For $\omega = (r_x, r_y) \in \mathcal{O}$, write $X' = r_x^\top X$, $Y' = r_y^\top Y$, $f(z) = \mathbb{E}(X' \mid Z = z)$, $g(z) = \mathbb{E}(Y' \mid Z = z)$, $\varepsilon = X' - f(Z)$, $\xi = Y' - g(Z)$, $u(z) = \mathbb{E}(\varepsilon^2 \mid Z = z)$, and $v(z) = \mathbb{E}(\xi^2 \mid Z = z)$, following Shah and Peters [34]. Let $\widehat{f}, \widehat{g}$ be the fitted conditional means and $p_n(\omega)$ the projected p-value of Section 3.2. The population target is $\theta(\omega) = \mathbb{E}(\varepsilon\xi)$. Conditional independence implies $\theta(\omega) = 0$, but not conversely.

Assumption E.18 (Projected GCM conditions)**.** The observations are i.i.d. For $\mathbb{P}_\mathcal{O}$-almost every ω, $\widehat{f}$ is fitted from $\{(X_i', Z_i)\}_{i=1}^n$ only and $\widehat{g}$ from $\{(Y_i', Z_i)\}_{i=1}^n$ only, with any auxiliary randomness independent of the data, $0 < \mathbb{E}(\varepsilon^2 \xi^2) < \infty$, and

$$A_f A_g = o_P(n^{-1}), \qquad B_f = o_P(1), \qquad B_g = o_P(1),$$

where

$$A_f = \frac{1}{n} \sum_{i=1}^n \{\widehat{f}(Z_i) - f(Z_i)\}^2, \qquad B_f = \frac{1}{n} \sum_{i=1}^n \{\widehat{f}(Z_i) - f(Z_i)\}^2 v(Z_i),$$

and A_g, B_g are defined symmetrically with $\widehat{g}$, g, and u.

These are the conditions of Shah and Peters [34, Theorem 6(i)] for each projected triplet. Fitting each regression only on its own response makes the cross terms $\{f(Z_i) - \widehat{f}(Z_i)\}\xi_i$ conditionally centered under the null, which permits in-sample fitting; the product condition lets each regression converge slowly provided the product of the two errors is $o_P(n^{-1})$.

E.4.2 Pointwise Type I error control

Theorem E.19 (Type I error control of RP-GCM). *Fix $k \geq 1$ and $P \in H_0$ satisfying Assumption E.18. If the projection pairs are drawn independently of the data, each with law $\mathbb{P}_{\mathcal{O}}$, then the RP-GCM decision $\phi_n = \mathbf{1}\{\min_{\ell \leq k} p_n(\omega_\ell) \leq \alpha/k\}$ satisfies $\limsup_{n\to\infty} \mathbb{P}_P(\phi_n = 1) \leq \alpha$. Neither the projection pairs nor the projected p-values need to be mutually independent.*

Proof. By Lemma E.1, $X' \perp\!\!\!\perp Y' \mid Z$ for every fixed ω. Under Assumption E.18, Shah and Peters [34, Theorem 6(i)] gives $T_n \Rightarrow N(0,1)$, so $\mathbb{P}_P\{p_n(\omega) \leq u\} \to u$ for every $u \in (0,1)$ and $\mathbb{P}_{\mathcal{O}}$-almost every ω. By bounded convergence, the same limit holds for an independent $\Omega \sim \mathbb{P}_{\mathcal{O}}$, and the union bound at $u = \alpha/k$ proves the claim. □

Uniform control. As in (1), the union bound gives

$$\mathbb{P}_P(\phi_n = 1) \leq \alpha + k\Delta_n(\alpha/k; P), \qquad \Delta_n(u; P) := \big[\mathbb{P}_P\{p_n(\Omega) \leq u\} - u\big]_+.$$

For a null class $\mathcal{P}_0$ and fixed k, uniform asymptotic control follows from $\sup_{P\in\mathcal{P}_0} \Delta_n(\alpha/k; P) \to 0$. Since $\Delta_n(u; P) \leq \int \sup_{P\in\mathcal{P}_0} [\mathbb{P}_P\{p_n(\omega) \leq u\} - u]_+ \, d\mathbb{P}_{\mathcal{O}}(\omega)$ and the integrand is bounded, it suffices that the uniform conditions of Shah and Peters [34, Theorem 6(ii)] hold for $\mathbb{P}_{\mathcal{O}}$-almost every ω. For a growing budget k_n, the bound requires $k_n \sup_{P\in\mathcal{P}_0} \Delta_n(\alpha/k_n; P) \to 0$. Pointwise validity alone gives neither guarantee.

E.4.3 Power under residual-covariance alternatives

Proposition E.20 (Power of RP-GCM). *Suppose $\mathbb{E}\|X\|^2, \mathbb{E}\|Y\|^2 < \infty$, and let*

$$\Sigma_{XY|Z} = \mathbb{E}\big[\{X - \mathbb{E}(X \mid Z)\}\{Y - \mathbb{E}(Y \mid Z)\}^\top\big].$$

Then $\theta(\omega) = r_x^\top \Sigma_{XY|Z} r_y$ for every $\omega \in \mathcal{O}$. If $\Sigma_{XY|Z} \neq 0$, then $\theta(\omega) \neq 0$ for $\mathbb{P}_{\mathcal{O}}$-almost every ω. If, in addition, the projection pairs are drawn independently of the data and $\mathbb{P}_P\{p_n(\omega) \leq \alpha/k\} \to 1$ for $\mathbb{P}_{\mathcal{O}}$-almost every ω with $\theta(\omega) \neq 0$, then the Type II error of RP-GCM tends to zero for every fixed $k \geq 1$.

Proof. Since $\varepsilon = r_x^\top\{X - \mathbb{E}(X \mid Z)\}$ and $\xi = r_y^\top\{Y - \mathbb{E}(Y \mid Z)\}$, we have $\theta(\omega) = \mathbb{E}(\varepsilon\xi) = r_x^\top \Sigma_{XY|Z} r_y$. If $\Sigma_{XY|Z} \neq 0$, the set of r_x with $\Sigma_{XY|Z}^\top r_x = 0$ is the intersection of the sphere with a proper subspace and is null. For any other r_x, the set of r_y with $\theta(\omega) = 0$ is a great subsphere and is null, so Fubini's theorem gives $\mathbb{P}_{\mathcal{O}}\{\theta(\Omega) = 0\} = 0$. For the second claim, condition on the projection pairs. With probability one, the first pair ω_1 has $\theta(\omega_1) \neq 0$ and lies outside the exceptional null set, and the aggregate rejects whenever its first projected test does. Hence the conditional Type II error is at most $\mathbb{P}_P\{p_n(\omega_1) > \alpha/k\} \to 0$, and bounded convergence over the projection pairs gives the claim. □

Asymptotically, a single projection suffices. In finite samples, k matters through the size of $|\theta(\Omega)|$: for uniform directions, $\mathbb{E}\,\theta(\Omega)^2 = \|\Sigma_{XY|Z}\|_F^2/(d_x d_y)$, so a single pair sees a diluted signal, and the maximum over k pairs selects larger values. Max-GCM targets $\max_{a,b} |(\Sigma_{XY|Z})_{ab}|$ instead, so RP-GCM tends to be favored when $\Sigma_{XY|Z}$ is dense or low rank rather than sparse. Scalar consistency is an additional requirement under the alternative; the sample-split power result of Shah and Peters [34, Theorem 8] does not by itself cover in-sample fitting. Dependence with $\Sigma_{XY|Z} = 0$, including conditional covariance that cancels across values of Z, is not targeted, since then $\theta(\omega) = 0$ for every ω.

Uniform power. On a class $\mathcal{P}_1$ of alternatives, with i.i.d. projection pairs, suppose that, uniformly in P, the pairs with $|\theta(\omega)| \geq \epsilon$ have $\mathbb{P}_{\mathcal{O}}$-probability at least $\pi_0 > 0$, and each projected test on these pairs fails to reject at level α/k with probability at most $b_n \to 0$, uniformly in P and ω. Conditioning on the projection pairs and testing the first such pair gives

$$\sup_{P\in\mathcal{P}_1} \mathbb{P}_P(\phi_n = 0) \leq (1 - \pi_0)^k + b_n.$$

Here k matters through π_0. Neither uniform bound is asserted for the OLS implementation without these additional scalar conditions.

E.4.4 Shared OLS implementation

Let $\mathbf{Z}_1 = [\mathbf{1}, \mathbf{Z}] \in \mathbb{R}^{n\times(d_z+1)}$ be the design matrix and $H = \mathbf{Z}_1\mathbf{Z}_1^+$ its orthogonal projection matrix. Multi-response OLS residuals $E_X = (I - H)\mathbf{X}$ and $E_Y = (I - H)\mathbf{Y}$ satisfy

$$E_X r_x = (I - H)(\mathbf{X}r_x), \qquad E_Y r_y = (I - H)(\mathbf{Y}r_y),$$

so projecting shared residuals equals fitting OLS after projection, and each projected fit uses only its own response and Z, as required by Assumption E.18. Standardizing the columns of $\mathbf{Z}$ leaves H unchanged. Our implementation uses these in-sample residuals, the empirical variance of R_i with divisor n, and two-sided normal p-values; numerically degenerate projections are assigned $p = 1$. Beyond the two regressions, projection and testing cost $O\{nk(d_x + d_y)\}$, instead of $2k$ separate nuisance fits. The identity holds for any fixed linear smoother, but not automatically for nonlinear or response-adaptively tuned learners.

Remark E.21 (Conditioning dimension under OLS). Suppose $\mathbb{E}(X \mid Z)$ and $\mathbb{E}(Y \mid Z)$ are linear in Z, and the conditional covariance matrices of X and Y given Z have operator norm at most C. Then $u, v \le C$ for every ω, and $\widehat{f}(Z_i) - f(Z_i) = (H\boldsymbol{\varepsilon})_i$ with $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, so

$$\mathbb{E}(A_f \mid \mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} H_{ii}u(Z_i) \le \frac{C(d_z + 1)}{n},$$

and likewise for A_g. Hence $A_f A_g = O_P(d_z^2/n^2)$ and $B_f, B_g = O_P(d_z/n)$, so $d_z = o(\sqrt{n})$ suffices for the regression conditions of Assumption E.18. If the conditional means are nonlinear, OLS residuals retain a function of Z and these conditions can fail under the null.

Finite-sample calibration. Remark E.21 is asymptotic. In finite samples, in-sample OLS removes $d_z + 1$ residual degrees of freedom, and under a homoscedastic Gaussian null $\mathrm{Var}(T_n) \approx \{1 - (d_z + 1)/n\}^{-1}$. The Bonferroni threshold α/k lies far in the tail, where this inflation is amplified: for $n = 500$ and $d_z = 50$, the factor 1.11 raises the per-projection exceedance probability at $\alpha/k = 5 \times 10^{-5}$ to about 1.2×10^{-4}. This is consistent with the small-sample null rejection rates in Appendix D.6.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contribution as a random-projection aggregation framework for CI testing with high-dimensional tested variables, and clarify the scope of the default CIT-based instantiation and its modular extensions. The stated theoretical and empirical claims match the results proved and evaluated in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the main limitations in Appendix B, including the scalar-Z scope of the default CIT-based analysis, fixed-k Type I guarantees, projection-dependent power, the practical nature of Beta-λ top-r calibration, and the regression-quality requirements of RP-GCM.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results state their assumptions explicitly and are cross-referenced. The main text provides the key statements and intuition, while complete proofs for Type I control, projection-capture power, growing-k power, top-r aggregation, and RP-GCM are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the data-generating mechanisms, experimental settings, baseline methods, evaluation protocols, and implementation details needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and experiment scripts are publicly available at `https://github.com/TJ-cyber-kkk/rp-ci`, with instructions for reproducing the experiments. The real datasets are public and are cited rather than redistributed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the main experimental settings, including sample sizes, dimensions, number of projections, significance level, Monte Carlo replications, baseline configurations, the image-valued benchmark, the known-DAG benchmark, and the Perturb-seq signature-skeleton setup, with further details in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Rejection rates are reported with exact binomial tests against the nominal level or 95% Wilson intervals; known-DAG comparisons use paired 95% confidence intervals, and the Perturb-seq consensus skeletons are compared with degree-preserving null graphs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on CPUs, mainly on a server with two Hygon C86 7470 processors; the paper reports algorithmic complexity and runtimes in Appendix D.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. It develops statistical methodology and uses public benchmark datasets without collecting new human-subject data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses broader impact in Appendix B: the main benefit is improved CI testing for high-dimensional scientific data analysis, while the main risk is over-interpretation of CI or skeleton-learning results as definitive causal conclusions outside the stated assumptions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, generative systems, scraped datasets, or other assets requiring special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets, including CIFAR-10 and the Norman Perturb-seq dataset, and cites the original sources. The released code downloads these datasets from their public sources rather than redistributing them. CIFAR-10 is used as distributed on its authors' page, which names no license. The raw Norman data are public in GEO under accession GSE133344, and we use the processed version distributed as scvi-tools example data. The baselines call causal-learn, released under the MIT license, and tigramite, released under GPL-3.0, as installed dependencies.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset is the RP-CIT code with experiment scripts, released under the MIT license at `https://github.com/TJ-cyber-kkk/rp-ci` and documented with instructions for reproducing the experiments.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or newly collected human-subject data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or newly collected human-subject data requiring IRB or equivalent approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology, experiments, or analysis.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.