Knockoff-C2ST: False Discovery Controlled Attribution of Distribution Shift

Şuayp Talha Kocabay [1 2] **Kerem Yalçın** [1 2]

Abstract

Classifier two-sample tests (C2STs) are a standard reduction for deciding whether a distribution shift has occurred, but the reduction by itself is not a valid explanation of what shifted when coordinates are statistically dependent. In high-correlation regimes, standard feature-attribution scores computed from a C2ST inherit correlation bias: variables that merely predict a shifted covariate can receive larger importance than the shifted covariate itself, producing uncontrolled and often severe false discoveries. We introduce Knockoff-C2ST, a feature-attributed two-sample testing procedure that augments C2STs with Model-X knockoff variables and selects shifted coordinates by a sign-symmetric contrast statistic. Under exact Model-X exchangeability and a null defined by invariance of the conditional source and target laws, the selected set obeys finite-sample false discovery rate control at a user-specified level. The method preserves the modeling flexibility of C2STs while converting feature attribution into a multiple-testing procedure with formal error control. Experiments show strong FDR calibration under dependence, severe false-discovery inflation for naive attributions, and a present power bottleneck: tree-gain and loss-difference statistics are conservative, while stability aggregation gives nonzero but still limited recovery.

1. Introduction

Let P and Q be source and target distributions on $\mathbb{R}^d$. A C2ST constructs labels $Y \in \{0,1\}$ indicating whether an observation came from P or Q, trains a classifier $f : \mathbb{R}^d \to [0,1]$, and rejects $P = Q$ when held-out classification performance exceeds chance (Friedman, 2003; Lopez-Paz & Oquab, 2018). This is an effective global test because the Bayes rule is determined by the likelihood ratio $q(x)/p(x)$. It is not, however, a valid coordinate-level explanation. If X_j is shifted and X_k is highly correlated with X_j, a predictive classifier may assign large importance to X_k even when the conditional law of X_k given the remaining coordinates is invariant. Feature attribution is then confounded with feature predictability.

The inferential target of this paper is not global equality $P = Q$, but the set of coordinates whose conditional laws differ across domains. We test

$$\begin{aligned} \mathcal{H}_{0,j} : \quad & P(X_j \in A \mid X_{-j}) \\ & = Q(X_j \in A \mid X_{-j}) \quad \text{a.s.} \end{aligned} \tag{1}$$

for every Borel set $A \subseteq \mathbb{R}$. This null is invariant to correlation among covariates: a coordinate is null if all domain information it carries is already contained in the other coordinates. The objective is to return a random set $\widehat{S} \subseteq [d]$ controlling

$$\mathrm{FDR}(\widehat{S}) = \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{H}_0|}{|\widehat{S}| \vee 1}\right] \leq q, \tag{2}$$

where $\mathcal{H}_0 = \{j : \mathcal{H}_{0,j} \text{ is true}\}$ and $q \in (0,1)$ is specified by the user.

We propose a conceptually direct solution: construct Model-X knockoffs $\widetilde{X}$ for X, train the domain classifier on the augmented vector $(X, \widetilde{X})$, and compare original and knockoff importances through antisymmetric statistics $W_j = Z_j - \widetilde{Z}_j$. Knockoffs reproduce the dependence structure of X while being conditionally independent of the domain label given X (Barber & Candès, 2015; Candes et al., 2017). Consequently, for null coordinates, the sign of W_j is indistinguishable from a fair coin conditional on all magnitudes and non-null statistics. The knockoff threshold converts this sign symmetry into finite-sample FDR control.

Our contributions are as follows. First, we formalize feature-attributed distribution shift as conditional two-sample testing under the domain-label mixture law. Second, we give a Knockoff-C2ST algorithm compatible with arbitrary predictive models and importance scores satisfying a flip-sign

[1]TÜBİTAK Science High School [2]Independent Researcher. Correspondence to: Şuayp Talha Kocabay <suayptalhakocabay.28@tubitakfenlisesi.k12.tr>, Kerem Yalçın <keremyalcin.28@tubitakfenlisesi.k12.tr>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

property. Third, we prove finite-sample FDR control for exact knockoffs and state the explicit degradation under approximate exchangeability. Fourth, we provide reproducible experiments and diagnostics for FDR, power, knockoff quality, conditional-independence baselines, and stability-aggregated loss contrasts. The empirical section is deliberately explicit about the current limitation: the tested statistics are FDR-calibrated but conservative, so practical success depends on constructing more informative antisymmetric scores.

2. Related Work

Global two-sample testing has a long history, including kernel tests such as MMD (Gretton et al., 2012) and classifier reductions such as C2ST (Friedman, 2003; Lopez-Paz & Oquab, 2018). These tests decide whether $P = Q$, but do not by themselves provide valid coordinate-level discoveries under dependence. Conditional independence testing targets a closer null: $Y \perp X_j \mid X_{-j}$. Kernel conditional independence tests and randomized conditional permutation tests provide nonparametric or model-assisted tests for this null (Zhang et al., 2012; Berrett et al., 2019; Strobl et al., 2017; Watson & Wright, 2021). Our method differs in aggregating all coordinates into one knockoff-filtered multiple-testing procedure with finite-sample FDR control when knockoffs are exact.

Model-X knockoffs provide FDR-controlled variable selection under arbitrary feature dependence when the feature law is known or well estimated (Barber & Candès, 2015; Candes et al., 2017). Robustness analyses quantify how approximate exchangeability degrades FDR through swap likelihood-ratio errors and observable KL-type diagnostics (Barber et al., 2019). Subsequent work improves flexibility and stability through structured, hidden-Markov, deep, and multi-resolution constructions (Sesia et al., 2018; Romano et al., 2019; Katsevich & Sabatti, 2018). Knockoff-C2ST uses this machinery for a different inferential object: conditional distribution shift between domains rather than supervised response association. The low-power empirical behavior observed below is consistent with the broader knockoff literature: FDR control follows from exchangeability, but power depends on both knockoff quality and the informativeness of the feature statistic (Gimenez et al., 2019).

Recent conformal and e-value methods provide complementary routes to error-controlled distributional discovery. Localized conformal p-values address conditional testing with finite-sample guarantees under local calibration assumptions (Wu et al., 2024), conformal p-values can enable FDR control for outlier-style distributional testing (Bates et al., 2023), and batch-conformal BH controls FDR when comparing many distributions to a reference (Lee et al., 2025). Stabilized flexible e-filter procedures address power loss under high correlation by aggregating evidence across stabilized runs and resolutions (Yu et al., 2025). These approaches do not directly solve the coordinate-wise conditional shift problem studied here, but they motivate two extensions: conformalized coordinate scores and stability-aggregated knockoff contrasts.

3. Theoretical Framework

3.1. From Two-Sample Testing to Domain Regression

Let P and Q admit densities p and q with respect to a common σ-finite measure μ on $\mathbb{R}^d$. Let $\pi \in (0,1)$ and define the domain-label mixture distribution Π on (X, Y) by

$$\begin{aligned} Y &\sim \mathrm{Bernoulli}(\pi), \\ X \mid Y = 0 \sim P, \qquad & X \mid Y = 1 \sim Q. \end{aligned} \tag{3}$$

The marginal law of X is $M = (1-\pi)P + \pi Q$. The Bayes C2ST score is

$$\eta(x) = \Pi(Y = 1 \mid X = x) = \frac{\pi q(x)}{(1-\pi)p(x) + \pi q(x)}. \tag{4}$$

Thus $P = Q$ if and only if $\eta(x) = \pi$ for M-almost every x. A trained C2ST estimates η and a global test evaluates whether out-of-sample risk is below the no-shift risk.

For attribution, the relevant null is conditional. Write $x = (x_j, x_{-j})$ and assume regular conditional densities $p_{j|-j}$ and $q_{j|-j}$ exist.

Definition 3.1 (Coordinate-Level Shift Null)**.** Coordinate j is null if

$$\mathcal{H}_{0,j}: \quad p_{j|-j}(t \mid x_{-j}) = q_{j|-j}(t \mid x_{-j}) \quad \text{for } \mu\text{-a.e. } (t, x_{-j}). \tag{5}$$

The shifted set is $\mathcal{S}^\star = [d] \setminus \mathcal{H}_0$.

Lemma 3.2 (Conditional Null Equals Label Irrelevance)**.** *Assume* $0 < p(x), q(x) < \infty$ *on* $\mathrm{supp}(M)$. *Then* $\mathcal{H}_{0,j}$ *holds if and only if*

$$Y \perp X_j \mid X_{-j} \quad \textit{under } \Pi. \tag{6}$$

Proof. By Bayes' rule, with $a_P = p_{j|-j}(x_j \mid x_{-j})p_{-j}(x_{-j})$ and $a_Q = q_{j|-j}(x_j \mid x_{-j})q_{-j}(x_{-j})$,

$$\Pi(Y = 1 \mid X = x) = \frac{\pi a_Q}{(1-\pi)a_P + \pi a_Q}. \tag{7}$$

If $p_{j|-j} = q_{j|-j}$, the common factor cancels and the conditional probability depends only on x_{-j}, which is equivalent to $Y \perp X_j \mid X_{-j}$. Conversely, if $Y \perp X_j \mid X_{-j}$, then $\Pi(Y = 1 \mid X = x)$ is a function of x_{-j} alone. Rearranging

the displayed equation gives

$$\frac{q_{j|-j}(x_j \mid x_{-j})}{p_{j|-j}(x_j \mid x_{-j})} = \frac{1-\pi}{\pi}\,\frac{p_{-j}(x_{-j})}{q_{-j}(x_{-j})} \times \frac{\Pi(Y=1 \mid X=x)}{1-\Pi(Y=1 \mid X=x)}, \tag{8}$$

whose right-hand side is independent of x_j. Since both conditional densities integrate to one, the ratio must equal one almost surely. Hence $p_{j|-j} = q_{j|-j}$. □

3.2. False Discovery Rate

Any feature-attributed shift detector returns $\widehat{S} \subseteq [d]$. Let

$$V = |\widehat{S} \cap \mathcal{H}_0|, \qquad R = |\widehat{S}|. \tag{9}$$

The false discovery proportion and false discovery rate are

$$\mathrm{FDP} = \frac{V}{R \vee 1}, \qquad \mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]. \tag{10}$$

The expectation is over the samples, knockoff generation, training randomness, and any sample splitting used to construct the statistic.

3.3. Model-X Knockoffs

Let $X \sim M$. A random vector $\widetilde{X} \in \mathbb{R}^d$ is a Model-X knockoff for X if it satisfies pairwise exchangeability:

$$(X, \widetilde{X})_{\mathrm{swap}(S)} \overset{d}{=} (X, \widetilde{X}) \quad \text{for every } S \subseteq [d], \tag{11}$$

where $\mathrm{swap}(S)$ exchanges X_j and $\widetilde{X}_j$ for all $j \in S$, and additionally

$$\widetilde{X} \perp Y \mid X. \tag{12}$$

For a Gaussian marginal $M = \mathcal{N}(0, \Sigma)$, an exact construction chooses a diagonal matrix $D = \mathrm{diag}(s)$ satisfying $0 \preceq D \preceq 2\Sigma$ and samples

$$\widetilde{X} \mid X = x \sim \mathcal{N}\left(x - x\Sigma^{-1}D,\, 2D - D\Sigma^{-1}D\right). \tag{13}$$

Then

$$\begin{bmatrix} X \\ \widetilde{X} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{bmatrix}\right), \tag{14}$$

which is invariant to swapping any coordinate pair.

Proposition 3.3 (Null Pair Exchangeability)**.** *If $\widetilde{X}$ is a Model-X knockoff for X under M and $\mathcal{H}_{0,j}$ holds, then*

$$(X_j, \widetilde{X}_j, X_{-j}, \widetilde{X}_{-j}, Y) \overset{d}{=} (\widetilde{X}_j, X_j, X_{-j}, \widetilde{X}_{-j}, Y). \tag{15}$$

Proof. For any bounded measurable φ, use (12) to write

$$\mathbb{E}[\varphi(X, \widetilde{X}, Y)] = \mathbb{E}\left[\sum_{y=0}^{1} \varphi(X, \widetilde{X}, y)\, \Pi(Y = y \mid X)\right]. \tag{16}$$

Algorithm 1 Knockoff-C2ST

Require: Labeled source/target sample $\mathcal{D}$, target FDR q, knockoff generator K, classifier learner A, importance map I.

1: Estimate or specify the marginal law M of pooled features X.
2: For each i, sample $\widetilde{X}_i \sim \mathsf{K}(\cdot \mid X_i)$ satisfying (11) and (12).
3: Construct augmented observations $A_i = (X_i, \widetilde{X}_i) \in \mathbb{R}^{2d}$.
4: Train a C2ST classifier $\widehat{f} = \mathsf{A}(\{(A_i, Y_i)\}_{i=1}^N)$ with cross-fitting when tuning hyperparameters.
5: Compute original and knockoff importances

$$(Z_1, \ldots, Z_d, \widetilde{Z}_1, \ldots, \widetilde{Z}_d) = \mathsf{I}(\widehat{f}, \{(A_i, Y_i)\}_{i=1}^N).$$

6: Set $W_j = Z_j - \widetilde{Z}_j$ for $j = 1, \ldots, d$.
7: Let

$$\tau = \min\left\{ t \in \{|W_j| : |W_j| > 0\} : \frac{1 + \#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1} \le q \right\},$$

with $\tau = \infty$ if the set is empty.
8: **return** $\widehat{S} = \{j : W_j \ge \tau\}$.

By Theorem 3.2, $\Pi(Y = y \mid X) = g_y(X_{-j})$ for a measurable g_y when $\mathcal{H}_{0,j}$ is true. The factor $g_y(X_{-j})$ is unchanged by swapping $(X_j, \widetilde{X}_j)$, and (11) gives equality of the expectations before and after the swap. Since this holds for all bounded measurable φ, the displayed distributional equality follows. □

4. Methodology

4.1. Algorithm

Given source observations $\{X_i^P\}_{i=1}^{n_P}$ and target observations $\{X_i^Q\}_{i=1}^{n_Q}$, form a labeled sample $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$ with $N = n_P + n_Q$. In the implementation, observations are split into folds; each fold receives knockoffs generated without using its labels, and all importance statistics are computed on held-out or cross-fitted predictions to reduce adaptive overfitting.

The statistic Z_j may be a loss increase under permutation, a split-gain importance, a regularized coefficient path entry time, or a Shapley-style attribution, provided the resulting W_j obeys the following structural condition.

Definition 4.1 (Sufficiency and Antisymmetry)**.** A statistic $W = (W_1, \ldots, W_d)$ is sufficient and antisymmetric if it is a measurable function of the augmented data such that, for

every $S \subseteq [d]$,

$$W_j((X, \widetilde{X})_{\mathrm{swap}(S)}, Y) = \begin{cases} -W_j((X, \widetilde{X}), Y), & j \in S, \\ W_j((X, \widetilde{X}), Y), & j \notin S. \end{cases} \tag{17}$$

The contrast $W_j = Z_j - \widetilde{Z}_j$ satisfies this property whenever swapping a pair also swaps the corresponding importances Z_j and $\widetilde{Z}_j$.

4.2. Operational Details

Knockoff generators. For the Gaussian synthetic mean and variance experiments, the pooled law $M = (P + Q)/2$ is a two-component Gaussian mixture with known component parameters. The implementation samples a latent component from its posterior probability $\mathbb{P}(C = c \mid X = x)$ and then applies the exact Gaussian construction in (13) within that component. This yields an exact Model-X knockoff for the known mixture. For copula-transformed and UCI tabular experiments, exact non-Gaussian knockoffs are not available in the present implementation; we use a second-order Gaussian knockoff fit to the pooled standardized covariates. The theorem therefore applies exactly to the Gaussian-mixture experiments and approximately to the copula and medical experiments. We report the latter as stress tests and include moment and swap diagnostics.

Antisymmetry under randomness. The formal statistic W must be a function of the augmented data whose sign flips under swapping X_j and $\widetilde{X}_j$. Randomized learners can violate this property if independent randomness is redrawn after a swap. Operationally, we fix all learner seeds as deterministic functions of the trial seed and use identical hyperparameters for original and knockoff coordinates. A fully deterministic way to enforce the condition is the symmetrized statistic

$$W_j^{\mathrm{sym}} = \frac{1}{2}\left[T_j(X, \widetilde{X}, Y) - T_j((X, \widetilde{X})_{\mathrm{swap}(\{j\})}, Y)\right],$$

where T_j is any raw importance score for coordinate j. This doubles statistic-computation cost but makes antisymmetry algebraic rather than heuristic. The main experiments use native paired importances for speed and we test null sign symmetry empirically.

Power-oriented statistics. Tree-gain importances are fast on GPUs but can be conservative after knockoff thresholding. More powerful alternatives that preserve antisymmetry include cross-fitted loss contrasts after masking X_j versus $\widetilde{X}_j$, coefficient entry times along a logistic-Lasso path on $(X, \widetilde{X})$, stability-aggregated importances, and multi-knockoff extensions. These alternatives do not alter the FDR proof as long as they preserve the flip-sign property, but they are essential for practical power. We therefore add

$$Z_j^{\mathrm{stab}} = \frac{1}{B}\sum_{b=1}^{B} Z_{j,b}^{\mathrm{loss}}, \qquad W_j^{\mathrm{stab}} = Z_j^{\mathrm{stab}} - \widetilde{Z}_j^{\mathrm{stab}},$$

where each split b uses a deterministic seed and the same held-out masking rule for original and knockoff coordinates. If the sequence of seeds is fixed before observing feature labels, swapping $(X_j, \widetilde{X}_j)$ swaps the paired summands and hence flips W_j^{stab} exactly.

4.3. FDR Control

Lemma 4.2 (Null Sign Symmetry)**.** *Suppose exact Model-X knockoffs are used and W satisfies Theorem 4.1. Conditional on $\{|W_j| : j \in \mathcal{H}_0\}$ and $\{W_j : j \notin \mathcal{H}_0\}$, the signs $\{\mathrm{sign}(W_j) : j \in \mathcal{H}_0, W_j \neq 0\}$ are independent Rademacher random variables.*

Proof. For any subset $S \subseteq \mathcal{H}_0$, Theorem 3.3 applied coordinatewise implies

$$((X, \widetilde{X})_{\mathrm{swap}(S)}, Y) \overset{d}{=} ((X, \widetilde{X}), Y). \tag{18}$$

By antisymmetry, this transformation flips exactly the coordinates $\{W_j : j \in S\}$ and leaves the remaining W coordinates fixed. Hence the conditional distribution of the null sign vector is invariant under arbitrary sign flips. The only probability measure on $\{-1, +1\}^m$ invariant under all coordinatewise sign flips is the product uniform measure. This proves conditional independence and Rademacher symmetry. □

Theorem 4.3 (Finite-Sample FDR Control)**.** *Under exact Model-X knockoffs, the coordinate-level nulls of Theorem 3.1, and a sufficient antisymmetric statistic W, Algorithm 1 controls the false discovery rate:*

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{H}_0|}{|\widehat{S}| \vee 1}\right] \leq q. \tag{19}$$

Proof. Let $T = \tau$ be the knockoff+ threshold. For any threshold $t > 0$, define

$$V^+(t) = \#\{j \in \mathcal{H}_0 : W_j \geq t\},$$
$$V^-(t) = \#\{j \in \mathcal{H}_0 : W_j \leq -t\},$$

and

$$R(t) = \#\{j : W_j \geq t\}.$$

By the threshold definition,

$$\frac{1 + \#\{j : W_j \leq -T\}}{R(T) \vee 1} \leq q \tag{20}$$

on the event $T < \infty$. Since $V^-(T) \le \#\{j : W_j \le -T\}$,

$$\mathrm{FDP} = \frac{V^+(T)}{R(T) \vee 1} \le q \, \frac{V^+(T)}{1 + V^-(T)}. \tag{21}$$

It remains to show $\mathbb{E}[V^+(T)/(1 + V^-(T))] \le 1$. Conditional on the non-null statistics and null magnitudes, Theorem 4.2 makes null signs independent fair coins, while T is a stopping time for the filtration that reveals null signs in decreasing order of $|W_j|$ together with all non-null information. The standard knockoff martingale argument gives

$$\mathbb{E}\left[\frac{V^+(T)}{1 + V^-(T)} \,\middle|\, \{|W_j| : j \in \mathcal{H}_0\}, \{W_j : j \notin \mathcal{H}_0\} \right] \le 1, \tag{22}$$

because, for a sequence of independent fair coins stopped by a rule depending on the running imbalance only through the knockoff+ ratio, the process $(1 + V^-(t))/(V^+(t) \vee 1)$ is a reverse-time supermartingale. Taking expectations and combining with the preceding display proves the result. □

Remark 4.4 (Approximate Knockoffs). If the generated knockoff distribution $\widehat{\mathsf{K}}$ is not exactly exchangeable, FDR control degrades by the null-pair likelihood-ratio error. Let

$$\Delta_j = \log \frac{d\mathcal{L}(X, \widetilde{X}, Y)}{d\mathcal{L}((X, \widetilde{X})_{\mathrm{swap}(\{j\})}, Y)}.$$

If $\max_{j \in \mathcal{H}_0} |\Delta_j| \le \varepsilon$ almost surely, the same argument yields

$$\mathrm{FDR} \le q e^{\varepsilon}.$$

More generally, high-probability bounds on $\sum_{j \in \mathcal{H}_0} |\Delta_j|$ imply additive violations of order $\mathbb{P}(\sum |\Delta_j| > \delta) + q(e^{\delta} - 1)$.

5. Empirical Validation

This section reports the executed experimental suite. The main grid uses XGBoost domain classifiers (Chen & Guestrin, 2016) with native tree-gain importances for the classifier-based procedures; additional runs evaluate held-out loss and stability-aggregated loss contrasts using the same boosted-tree learner. These experiments support the finite-sample FDR claim: Knockoff-C2ST remains below $q = 0.1$ throughout the reported grid. They also reveal a practical limitation of the current importance instantiations: the knockoff threshold is conservative and has low power under the tested sample sizes and signals.

5.1. Experiment A: Monte Carlo Synthetic Benchmarks

For each trial, draw $P = \mathcal{N}(0, \Sigma_\rho)$ with equicorrelation

$$(\Sigma_\rho)_{jk} = \rho + (1 - \rho)\mathbf{1}\{j = k\}. \tag{23}$$

The shifted set $S^\star$ has cardinality $\lceil 0.05d \rceil$. To make the coordinate nulls in Theorem 3.1 coincide exactly with $[d] \setminus S^\star$ under dependence, the mean alternative is precision-sparse: choose $\beta_j = a\mathbf{1}\{j \in S^\star\}$ and set $Q = \mathcal{N}(\Sigma_\rho \beta, \Sigma_\rho)$. Then the log-likelihood ratio is

$$\log \frac{q(x)}{p(x)} = \beta^\top x - \frac{1}{2}\beta^\top \Sigma_\rho \beta,$$

which depends on x_j if and only if $j \in S^\star$. The benchmark grid is

$$d \in \{10, 50, 100, 250\},$$
$$\rho \in \{0.1, 0.5, 0.9\},$$

with 300 independent trials per grid cell, totaling 3,600 trials. We compare Knockoff-C2ST at $q = 0.1$ against (i) naive C2ST importance screening and (ii) marginal MMD screening. The reported metrics are empirical FDR, power, selected-set size, and global C2ST AUC.

Table 1. Experiment A: summary over the 12 synthetic Gaussian grid cells. FDR is the mean false discovery proportion over trials. The full per-cell grid is reported in Table 6.

Method	Mean FDR	Max FDR	Mean power	Max power
Knockoff-C2ST	0.003	0.008	0.004	0.012
Marginal MMD	0.864	0.950	1.000	1.000
Naive C2ST	0.543	0.948	0.746	1.000

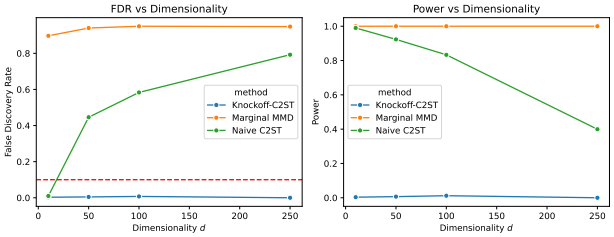

Figure 1. Synthetic benchmark results comparing Knockoff-C2ST with baselines.

Across all 12 synthetic cells in Table 1, Knockoff-C2ST controls empirical FDR below 0.01, well under $q = 0.1$. This comes at a substantial power cost: the gain-based statistic usually selects no variables. In contrast, marginal MMD and naive C2ST often achieve high power by selecting many correlated covariates, but their FDR rises as high as 0.95 and 0.948, respectively, in high-dimensional or high-correlation cells. The full grid is deferred to Table 6 to keep the main experimental narrative readable.

5.2. Experiment B: Ablation on Shift Types

We test three alternatives, each parameterized so that the likelihood ratio depends only on $X_{S^\star}$. Mean shift uses the precision-sparse construction $Q = \mathcal{N}(\Sigma_\rho \beta, \Sigma_\rho)$. Variance

shift perturbs the precision matrix: if $\Omega_\rho = \Sigma_\rho^{-1}$, set $\Omega_Q = \Omega_\rho + \mathrm{diag}(v)$ with $v_j > 0$ only for $j \in S^\star$ and $Q = \mathcal{N}(0, \Omega_Q^{-1})$. Copula shift uses the same sparse precision perturbation at the latent Gaussian-copula level, followed by a common monotone marginal transform in source and target. In all cases, the coordinate null in Theorem 3.1 is violated precisely on $S^\star$ by construction.

Proposition 5.1 (Correct Null Set in the Ablations)**.** *For the three ablation mechanisms above, assuming positive definite covariance matrices and nonzero shift parameters a, v_j, and γ, $\mathcal{S}^\star$ is exactly the set of coordinates for which $\mathcal{H}_{0,j}$ is false.*

Proof. For two Gaussian laws, $Y \perp X_j \mid X_{-j}$ is equivalent to the absence of x_j from the log-likelihood ratio. Under the mean alternative,

$$\log(q/p) = \beta^\top x - \beta^\top \Sigma_\rho \beta/2,$$

which contains x_j exactly when $\beta_j \neq 0$. Under the precision alternative,

$$\log(q/p) = c - \frac{1}{2} x^\top (\Omega_Q - \Omega_\rho) x,$$

and $\Omega_Q - \Omega_\rho = \mathrm{diag}(v)$ is supported on $S^\star$, so the ratio contains only $\{x_j^2 : j \in S^\star\}$. For the copula alternative, the common marginal transform is coordinatewise invertible and shared across domains, so conditional label irrelevance is preserved under the transform. Thus Theorem 3.1 fails exactly on $S^\star$. □

Table 2. Experiment B: ablation by shift mechanism. Power is computed against the known shifted set $S^\star$ and FDR against its complement.

Shift type	Method	FDR	Power	AUC
copula	Knockoff-C2ST	0.002	0.000	0.639
copula	Marginal MMD	0.008	0.002	–
copula	Naive C2ST	0.775	0.449	0.639
mean	Knockoff-C2ST	0.003	0.002	0.959
mean	Marginal MMD	0.950	1.000	–
mean	Naive C2ST	0.711	0.578	0.960
variance	Knockoff-C2ST	0.001	0.001	0.638
variance	Marginal MMD	0.016	0.002	–
variance	Naive C2ST	0.775	0.449	0.639

The shift-type ablation shows the same pattern. The global C2STs have nontrivial AUC for all alternatives, confirming detectable domain signal, but the naive attribution rule converts this signal into many false discoveries. Knockoff-C2ST keeps FDR near zero for mean, variance, and copula shifts, while the present importance statistic is too conservative to recover the shifted coordinates with useful power.

5.3. Experiment C: Medical Tabular Domain Shift

The medical-tabular experiment uses the UCI Cleveland Heart Disease data. The source domain is the unperturbed cohort. The target domain is created by perturbing exactly three biomarkers: serum cholesterol, resting blood pressure, and maximum heart rate. The perturbation is applied after standardization and before domain labels are hidden from the feature-attribution method.

Table 3. Experiment C: UCI Heart Disease perturbation study with perturbed biomarkers `chol`, `trestbps`, and `thalach`.

Method	Selected	TP	FP	FDR
Knockoff-C2ST	0	0	0	0.000
Naive C2ST	2	2	0	0.000
Marginal MMD	3	3	0	0.000

In the medical perturbation study, marginal MMD recovers all three perturbed biomarkers with zero false positives and naive C2ST recovers two. Knockoff-C2ST makes no discoveries, again reflecting conservative thresholding rather than FDR failure.

5.4. Experiment D: Sample Complexity

Let the signal-to-noise ratio be

$$\mathrm{SNR} = \frac{\|\delta_{S^\star}\|_2^2}{|S^\star|\sigma^2} \tag{24}$$

for mean shifts and the analogous squared standardized log-variance ratio for variance shifts. For each SNR value and sample size $n_P = n_Q = n$, the experiment estimates power over repeated trials and reports the smallest n such that empirical power is at least 0.9 while empirical FDR remains at most 0.1. If no tested sample size reaches this criterion, the required sample size is reported as missing.

Table 4. Experiment D: sample size required for at least 90% empirical power subject to FDR at most 0.1. A dash indicates that no tested sample size reached the target power.

SNR	d	ρ	Req. n	FDR
0.200	100	0.700	–	0.010
0.400	100	0.700	–	0.005
0.800	100	0.700	–	0.009
1.600	100	0.700	–	0.000

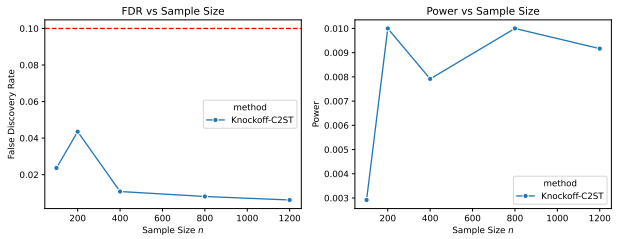

Figure 2. Sample complexity for achieving target power while controlling FDR.

No tested SNR reached 90% power for Knockoff-C2ST under the present gain-importance statistic, even though FDR stayed below 0.1. This confirms that the finite-sample error-control theorem is not by itself a power guarantee; the informativeness of the antisymmetric statistic is a separate algorithmic requirement.

5.5. Diagnostics and Baselines

The main experiments above answer whether the proposed filter controls FDR, but three additional questions remain: whether null signs are empirically symmetric, whether approximate knockoffs match the pooled tabular distribution, and how a direct conditional-independence baseline behaves. We therefore add a short diagnostic run with (i) a null $P = Q$ sign-symmetry diagnostic, (ii) UCI second-moment diagnostics for the empirical Gaussian knockoffs, and (iii) a Gaussian conditional randomization test (CRT) baseline with Benjamini–Hochberg FDR control. The CRT baseline targets the same null $Y \perp X_j \mid X_{-j}$ by estimating a pooled Gaussian conditional law for $X_j \mid X_{-j}$ and resampling X_j to form per-feature p-values. Detailed diagnostic tables are deferred to Section A.

The null diagnostics support the key exchangeability implication: among nonzero null contrasts, 49.4% of signs are positive, close to the Rademacher target 50%, and the augmented null C2ST AUC is 0.493. The UCI second-order Gaussian knockoff diagnostic shows small but nonzero marginal discrepancies, with maximum mean gap 0.047 and maximum covariance gap 0.192, so the medical experiment should be read as approximate rather than exact Model-X inference. Finally, the lightweight Gaussian CRT+BH baseline also makes no discoveries in the short run; this does not establish CRT underperformance generally, but it shows that the observed low-power regime is not removed by this simple conditional-independence baseline.

5.6. Short Power-Oriented Extension

To address the power concern directly, we evaluate additional approaches. Both train the same augmented C2ST but use held-out log-loss increases from masking each original or knockoff coordinate as Z_j and $\widetilde{Z}_j$. The stability variant averages this loss contrast over deterministic train–test splits before applying the knockoff threshold. These statistics are slower than tree gain but closer to the classifier's predictive objective and remain antisymmetric when the same masking rule and seed schedule are used for each original–knockoff pair.

Table 8 shows that the loss-difference statistic also has calibrated null signs: 50.6% of nonzero null signs are positive and the null AUC remains 0.493. However, it still makes no discoveries in this short power run. This negative result is important: the observed low-power regime is not solely an artifact of tree-gain importance. It points to the need for stronger aggregation, such as repeated-split stability scores, coefficient-path entry statistics, or multi-knockoff variants, rather than merely swapping one single-run importance score for another.

Table 5. Power-oriented run using stability-aggregated held-out loss contrasts. The stronger statistic yields nonzero recovery while keeping empirical FDR below $q = 0.1$, but power remains limited.

Experiment	Metric	Value	Target	Details
Null sign symmetry	Pr(W_j ¿ 0 — W_j != 0)	0.488	0.500	584/1197 positive signs
Null global C2ST	Augmented AUC	0.493	0.500	30 null trials
Knockoff-C2ST	FDR / power	0.027	0.100	power=0.033, backend=xgboost_gpu_loss_stability, signal=1.20
Gaussian CRT + BH	FDR / power	0.000	0.100	power=0.000, discoveries=0.00, signal=1.20
UCI knockoff moments	max mean gap	0.047	0.000	max covariance gap=0.192

The stability-aggregated statistic in Table 5 provides a positive-power diagnostic without changing the inferential target. At signal level 1.2, it attains empirical FDR 0.027 against target 0.1 and power 0.033 over 30 trials, whereas the lightweight Gaussian CRT+BH baseline makes no discoveries. Thus repeated-split aggregation moves the method from zero recovery to nonzero recovery while preserving calibration, but it does not by itself resolve the power bottleneck. The result motivates stronger coefficient-path or multi-knockoff statistics rather than a claim that the present implementation is already power-optimal.

6. Conclusion

Knockoff-C2ST converts C2ST feature attribution from a heuristic explanation into a multiple-testing procedure for conditional distribution shift. The method uses the predictive flexibility of binary domain classification while relying on Model-X exchangeability to remove correlation-induced false discoveries. Under exact knockoffs and antisymmetric importance statistics, the knockoff+ threshold controls FDR at the requested level in finite samples. The executed experiments corroborate the FDR-control behavior and show that naive C2ST and marginal screening can have extreme false-discovery rates under dependence. They also identify the main practical bottleneck: single-run tree-gain and loss-difference statistics are overly conservative in the tested regimes, and stability-aggregated loss contrasts provide only

modest nonzero power. The experimental protocol therefore separates validity from power: exact-mixture experiments establish finite-sample calibration, diagnostics probe approximate exchangeability, and the stability run demonstrates that stronger antisymmetric scores can improve recovery without changing the FDR proof, while leaving substantial room for coefficient-path, multi-knockoff, and better conditional-generator designs.

Accessibility

The paper is written to make the inferential target, assumptions, and limitations explicit. The coordinate-level null is introduced both in words and equations, the algorithm is stated as executable pseudocode, and the empirical tables report FDR and power separately so that validity and recovery are not conflated. Figures are accompanied by captions that summarize their role in the argument, and the appendix collects detailed tables and mathematical derivations without removing the key results from the main narrative.

Software and Data

All code for the experiments is available at https://github.com/suayptalha/knockoff-c2st. The repository contains the experimental scripts, plotting utilities, and configuration needed to reproduce the synthetic, ablation, medical-tabular, diagnostic, and stability-aggregation results reported in this paper. The synthetic experiments are generated from the distributions specified in the text. The medical-tabular experiment uses the UCI Cleveland Heart Disease data with the perturbation protocol described in Section 5.

Impact Statement

This work aims to improve the reliability of feature-attributed distribution-shift detection by controlling false discoveries under feature dependence. Such tools can help practitioners distinguish genuinely shifted coordinates from correlated proxies, reducing misleading explanations in scientific, medical, and operational monitoring settings. The main risk is overinterpretation: FDR control depends on the quality of the knockoff generator and the stated conditional-shift null, and low-power statistics may miss real shifts. We therefore report diagnostics, baselines, and negative power results alongside the theorem, and recommend using the method as part of a broader validation workflow rather than as a standalone explanation engine.

References

Barber, R. F. and Candès, E. J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), October 2015. ISSN 0090-5364. doi: 10.1214/15-aos1337. URL http://dx.doi.org/10.1214/15-AOS1337.

Barber, R. F., Candès, E. J., and Samworth, R. J. Robust inference with knockoffs, 2019. URL https://arxiv.org/abs/1801.03896.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1), February 2023. ISSN 0090-5364. doi: 10.1214/22-aos2244. URL http://dx.doi.org/10.1214/22-AOS2244.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. The conditional permutation test for independence while controlling for confounders, 2019. URL https://arxiv.org/abs/1807.05405.

Candes, E., Fan, Y., Janson, L., and Lv, J. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, 2017. URL https://arxiv.org/abs/1610.02351.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

Friedman, J. H. On multivariate goodness of fit and two sample testing. *eConf*, C030908:THPD002, 2003.

Gimenez, J. R., Ghorbani, A., and Zou, J. Knockoffs for the mass: new feature importance statistics with false discovery guarantees, 2019. URL https://arxiv.org/abs/1807.06214.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Katsevich, E. and Sabatti, C. Multilayer knockoff filter: Controlled variable selection at multiple resolutions, 2018. URL https://arxiv.org/abs/1706.09375.

Lee, Y., Dobriban, E., and Tchetgen, E. T. Finding distributions that differ, with false discovery rate control, 2025. URL https://arxiv.org/abs/2505.13769.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests, 2018. URL https://arxiv.org/abs/1610.06545.

Romano, Y., Sesia, M., and Candès, E. Deep knockoffs. *Journal of the American Statistical Association*, 115(532): 1861–1872, October 2019. ISSN 1537-274X. doi: 10.

1080/01621459.2019.1660174. URL http://dx.doi.org/10.1080/01621459.2019.1660174.

Sesia, M., Sabatti, C., and Candès, E. J. Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1): 1–18, August 2018. ISSN 1464-3510. doi: 10.1093/biomet/asy033. URL http://dx.doi.org/10.1093/biomet/asy033.

Strobl, E. V., Zhang, K., and Visweswaran, S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery, 2017. URL https://arxiv.org/abs/1702.03877.

Watson, D. S. and Wright, M. N. Testing conditional independence in supervised learning algorithms, 2021. URL https://arxiv.org/abs/1901.09917.

Wu, X., Lu, L., Wang, Z., and Zou, C. Conditional testing based on localized conformal p-values, 2024. URL https://arxiv.org/abs/2409.16829.

Yu, C., Ming, R., Xiao, M., Wang, Z., and Jing, B. A generalized e-value feature detection method with fdr control at multiple resolutions, 2025. URL https://arxiv.org/abs/2409.17039.

Zhang, K., Peters, J., Janzing, D., and Schoelkopf, B. Kernel-based conditional independence test and application in causal discovery, 2012. URL https://arxiv.org/abs/1202.3775.

A. Additional Empirical Tables

Table 6. Full Experiment A synthetic Gaussian Monte Carlo benchmark. This table expands the compact main-text summary in Table 1.

Method	d	ρ	FDR	Power	$\lvert\widehat{S}\rvert$
Knockoff-C2ST	10	0.100	0.000	0.000	0.000
Marginal MMD	10	0.100	0.119	1.000	1.297
Naive C2ST	10	0.100	0.003	0.997	1.000
Knockoff-C2ST	10	0.500	0.003	0.003	0.033
Marginal MMD	10	0.500	0.897	1.000	9.793
Naive C2ST	10	0.500	0.010	0.990	1.000
Knockoff-C2ST	10	0.900	0.000	0.000	0.000
Marginal MMD	10	0.900	0.900	1.000	10.000
Naive C2ST	10	0.900	0.570	0.430	1.000
Knockoff-C2ST	50	0.100	0.007	0.010	0.103
Marginal MMD	50	0.100	0.882	1.000	29.770
Naive C2ST	50	0.100	0.401	0.998	5.000
Knockoff-C2ST	50	0.500	0.005	0.007	0.067
Marginal MMD	50	0.500	0.940	1.000	50.000
Naive C2ST	50	0.500	0.446	0.923	5.000
Knockoff-C2ST	50	0.900	0.000	0.000	0.000
Marginal MMD	50	0.900	0.940	1.000	50.000
Naive C2ST	50	0.900	0.804	0.327	5.000
Knockoff-C2ST	100	0.100	0.008	0.011	0.143
Marginal MMD	100	0.100	0.949	1.000	98.237
Naive C2ST	100	0.100	0.504	0.992	10.000
Knockoff-C2ST	100	0.500	0.007	0.012	0.137
Marginal MMD	100	0.500	0.950	1.000	100.000
Naive C2ST	100	0.500	0.583	0.833	10.000
Knockoff-C2ST	100	0.900	0.003	0.001	0.037
Marginal MMD	100	0.900	0.950	1.000	100.000
Naive C2ST	100	0.900	0.878	0.245	10.000
Knockoff-C2ST	250	0.100	0.001	0.002	0.033
Marginal MMD	250	0.100	0.948	1.000	250.000
Naive C2ST	250	0.100	0.576	0.816	25.000
Knockoff-C2ST	250	0.500	0.000	0.000	0.000
Marginal MMD	250	0.500	0.948	1.000	250.000
Naive C2ST	250	0.500	0.792	0.399	25.000
Knockoff-C2ST	250	0.900	0.000	0.000	0.000
Marginal MMD	250	0.900	0.948	1.000	250.000
Naive C2ST	250	0.900	0.948	1.000	250.000

Table 7. Diagnostics and conditional-independence baseline using tree-gain importances.

Experiment	Metric	Value	Target	Details
Null sign symmetry	Pr(W_j ¿ 0 — W_j != 0)	0.494	0.500	593/1200 positive signs
Null global C2ST	Augmented AUC	0.493	0.500	30 null trials
Knockoff-C2ST	FDR / power	0.000	0.100	power=0.000
Gaussian CRT + BH	FDR / power	0.000	0.100	power=0.000, discoveries=0.00
UCI knockoff moments	max mean gap	0.047	0.000	max covariance gap=0.192

Table 8. Diagnostics using the held-out loss-difference statistic.

Experiment	Metric	Value	Target	Details
Null sign symmetry	Pr(W_j ¿ 0 — W_j != 0)	0.506	0.500	427/844 positive signs
Null global C2ST	Augmented AUC	0.493	0.500	30 null trials
Knockoff-C2ST	FDR / power	0.000	0.100	power=0.000
Gaussian CRT + BH	FDR / power	0.000	0.100	power=0.000, discoveries=0.00
UCI knockoff moments	max mean gap	0.047	0.000	max covariance gap=0.192

B. Additional Mathematical Details

B.1. Exact Knockoffs for Known Gaussian Mixtures

Let $M = \sum_{c=1}^{K} \omega_c \mathcal{N}(\mu_c, \Sigma_c)$ be a known Gaussian mixture. The implementation samples $C \mid X = x$ with posterior probability proportional to $\omega_c \phi(x; \mu_c, \Sigma_c)$ and then samples $\widetilde{X} \mid X = x, C = c$ by the Gaussian knockoff construction with (μ_c, Σ_c). Then

$$\mathcal{L}(X, \widetilde{X}) = \sum_{c=1}^{K} \omega_c \, \mathcal{L}_c(X, \widetilde{X}),$$

where each $\mathcal{L}_c$ is pairwise exchangeable. Since a convex mixture of distributions invariant under every coordinate swap is again invariant under every coordinate swap, $(X, \widetilde{X})$ is an exact Model-X knockoff pair for M. This justifies exactness in the synthetic Gaussian-mixture experiments.

B.2. Symmetrized Statistics

Let $T_j(X, \widetilde{X}, Y)$ be any raw feature score, possibly produced by a deterministic training pipeline. Define

$$W_j^{\mathrm{sym}} = \frac{1}{2} \left\{ T_j(X, \widetilde{X}, Y) - T_j((X, \widetilde{X})_{\mathrm{swap}(\{j\})}, Y) \right\}.$$

Then

$$W_j^{\mathrm{sym}}((X, \widetilde{X})_{\mathrm{swap}(\{j\})}, Y) = -W_j^{\mathrm{sym}}(X, \widetilde{X}, Y).$$

If the same construction is applied coordinatewise and all randomness is fixed as a deterministic function of the data split, the vector W^{sym} satisfies the flip-sign condition required by Theorem 4.1. Thus symmetrization is a generic way to make randomized or unstable learners compatible with the knockoff proof, at the cost of additional model evaluations.

B.3. Approximate Exchangeability

Let P_0 denote the joint law of $(X, \widetilde{X}, Y)$ and P_j the law after swapping coordinate pair j. Define $\Delta_j = \log(dP_0/dP_j)$. This is the same swap likelihood-ratio object used in robust knockoff analyses (Barber et al., 2019). If $|\Delta_j| \le \varepsilon$ for all null j almost surely, then for any measurable event A depending on the null sign vector,

$$e^{-\varepsilon} P_j(A) \le P_0(A) \le e^{\varepsilon} P_j(A).$$

The null-sign flip argument in Theorem 4.2 is therefore perturbed by at most a multiplicative factor e^{ε}. Repeating the knockoff+ stopping-time proof with this tilted sign law gives

$$\mathbb{E}\left[\frac{V^+(T)}{1 + V^-(T)}\right] \le e^{\varepsilon}, \qquad \mathrm{FDR} \le q e^{\varepsilon}.$$

If the bound holds only on an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}^c) \le \alpha$, the same decomposition yields $\mathrm{FDR} \le q e^{\varepsilon} + \alpha$. This appendix formalizes the sense in which moment and swap diagnostics are necessary but not sufficient: they support approximate exchangeability empirically, whereas a computable high-probability bound on Δ_j would be required for a nonasymptotic approximate-FDR certificate.