Learning Kernelized Hypothesis for Hidden Confounder Detection

Yikai Chen[1] , **Haotian Wang**[1*] , **Yunxin Mao**[1] , **Xinpeng Lv**[1] , **Shaowu Yang**[1] , **Kun Kuang**[2] , **Xinwang Liu**[1] , **Wenjing Yang**[1]

[1]National University of Defense Technology, Changsha, China
[2]Zhejiang University, Hangzhou, China
{chenyikai1998, wanghaotian13, maoyunxin, lvxinpeng, shaowu.yang}@nudt.edu.cn, kunkuang@zju.edu.cn, {xinwangliu, wenjing.yang}@nudt.edu.cn

Abstract

Detecting hidden confounding is crucial for reliable causal analysis from observational data, directly determining which downstream causal inference method to be deployed. Inspired by the theory of higher-order regression, recent sample-efficient hypothesis testing strategies overcome the restrictive requirement of multiple heterogeneous data environment. Despite their progress on single-environment confounder detection, such methods suffer from intrinsic flaws that the structural functions of the causal models should be specified in prior (linear or specific kernel functions). By contrast, real-world data acquisition exhibits diverse, unknown forms of structural functions, imposing an important but challenging gap between theories of higher-order regressions to practical confounder detection. In this paper, we contribute a Bi-level Kernel Confounder Detection (BiKCD) framework by learning adaptive kernelized structural space of structural functions. Subsequently, our BiKCD constructs hypothesis testing by comparing coefficients from the higher-order regression and the classical ordinary least squares in learned kernelized space. Finally, the hypothesis is calibrated to ensure valid inference under adaptivity. Theoretically, we establish an oracle-type risk bound for the selected structural space over a candidate kernel family, with the Type-I error control for the downstream test. Extensive experiments on synthetic and real-world datasets demonstrate the effectiveness of the proposed BiKCD.

1 Introduction

Causal effect estimation plays a central role in many high-stakes domains, including healthcare, public policy, and finance [Prosperi *et al.*, 2020; Dahabreh and Bibbins-Domingo, 2024; Matthay and Glymour, 2022; Kühne *et al.*, 2022; Vanderschueren, 2024]. Although randomized controlled trials provide the most reliable evidence, they are often infeasible due to ethical, operational, or cost constraints. Consequently, observational studies remain a primary source of evidence in practice [Pearl, 2009; Listl *et al.*, 2016]. However, credible causal conclusions from observational data typically hinge on the unconfoundedness assumption, namely that all variables influencing both treatment and outcome are observed and properly adjusted for. When this assumption fails due to hidden confounders, downstream causal estimators can be systematically biased, even when equipped with powerful nonlinear prediction models [Angrist *et al.*, 1996]. This makes confounder detection and falsification procedures a critical diagnostic step that can guide whether, and how, one should proceed with causal effect estimation [Lee and Lemieux, 2010].

To detect hidden confounding, a dominant line of recent progress leverages heterogeneity across multiple environments by observing changes in causal mechanisms to create testable implications [Karlsson and Krijthe, 2023; Mameche *et al.*, 2024; Reddy and N Balasubramanian, 2024]. While powerful in principle, these methods impose stringent data requirements: sufficiently diverse environments must be observed, and the underlying causal structure must satisfy invariance or mechanism-independence conditions across environments. In many real-world applications, however, it is hard to collect multiple observed environments with sufficiently shift-based diagnostics. By contrast, only a single observational dataset is available in most practical cases. Therefore, this gap motivates effective confounder detection methods towards the single-environment setting.

To detect hidden confounders in a single data environment, recent regression-based approaches provide a promising route by following a regression-comparison framework. To be specific, ROCD compares ordinary least squares with a higher-order regression analogue to detect violations of linear causal model assumptions, using their discrepancy as a diagnostic signal [Schultheiss *et al.*, 2024]. Subsequently, KRCD extends the comparison to nonlinear settings by moving from a linear function class to a kernelized space by contrasting a standard kernel least-squares regression with a norm-weighted higher-order kernel regression [Chen *et al.*, 2026]. Consequently, more practical confounder detection under nonlinear mechanisms are facilitated.

In this paper, we aim to investigate *practical* hidden confounder detection in a *single observational environment*. Unfortunately, when bridging existing ROCD and KRCD to

*Corresponding Author.

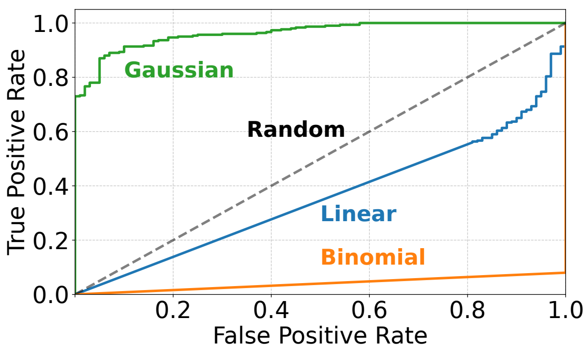


Figure 1: KRCD ROC curves on an exponential synthetic benchmark with different fixed kernel classes.

practical, single-environment data, they still suffer from the reliance on the specification of structural-function in prior. ROCD assumes linear structural functions, and KRCD must specify the non-linear kernel function. In contrast, structural functions of the true data-generating mechanism are unknown, spanning a diversity of nonlinear forms. As a result, such fixed specifications of structural functions leads to degraded detection performance, as it is difficult to distinguish genuine confounding signals from possible structural mismatch of structural functions. For example, on an exponential synthetic data environment, KRCD yields a strong ROC curve only when the kernel class is well aligned with the mechanism (Figure 1). Thus, a confounder detection framework with adaptive structural space of structural functions for single-environment confounder detection is motivated.

To this end, we propose a learnable and flexible *Bi-level Kernel Confounder Detection (BiKCD)* approach through a concise three-part pipeline. BiKCD replaces a fixed, user-specified structural function class with a data-driven kernelized structural space learned via a bi-level optimization scheme, where the outer level learns an effective kernel structure and the inner regression admits closed-form solutions. Given this learned structural space, BiKCD performs confounder detection in the induced kernel space by contrasting regression coefficients from a higher-order (norm-weighted) regression with those from the classical least-squares regression, using their discrepancy as the diagnostic signal. The resulting hypothesis test is finally calibrated under adaptivity using selection-test separation (sample splitting or cross-fitting), producing valid p-values with controlled Type-I error while improving robustness to structural misspecification in single-environment settings. The main contributions of this paper can be summarized as follows:

- We propose BiKCD, a bi-level kernel confounder detection framework that learns adaptive kernelized structural spaces for practical hidden confounder detection in single-environment observational data with unknown nonlinear mechanisms.
- Theoretically, we establish an oracle-type risk bound for the selected structural space over a candidate kernel family and guarantee Type-I error control for the downstream hypothesis test under adaptivity.
- Empirically, we demonstrate through extensive experiments on synthetic and real-world datasets that BiKCD effectively detects hidden confounders and is robust to structural misspecification.

2 Related Work

2.1 Detection of Unobserved Confounders in Single-Environment Settings

Traditional diagnostics for hidden confounding include instrumental variable (IV) regression [Angrist *et al.*, 1996] and specification tests based on estimator disagreement, such as the Durbin–Wu–Hausman test [Hausman, 1978]. These methods typically require external instruments or alternative estimators whose consistency hinges on different assumptions, and they are often studied alongside practical IV constructions and variants in applied settings [Wang *et al.*, 2022; Wang *et al.*, 2026; Kou *et al.*, 2025]. A more recent line replaces instruments with moment-based contrasts and regression-oriented model checks: ROCD [Schultheiss *et al.*, 2024] builds on reweighting views of model checking [Buja *et al.*, 2019a; Buja *et al.*, 2019b] and compares ordinary least squares to a higher-order least-squares analogue in a resampling-free manner [Wang *et al.*, 2023a]. For high-dimensional linear models, spectral approaches detect confounding by examining regression-induced associations and covariance perturbations [Liu and Chan, 2018; Janzing and Schölkopf, 2018]. Beyond linearity, latent-variable structure learning based on higher-order information has been explored under identifiable assumptions [Cai *et al.*, 2023; Chen *et al.*, 2024], while other approaches address nonlinear confounding through strong structural requirements (e.g., nonparametric IV [Newey and Powell, 2003]) or specific non-Gaussian generative models (e.g., LiNGAM [Shimizu *et al.*, 2006]). Within kernel methods, KRCD [Chen *et al.*, 2026] introduced an RKHS diagnostic that embeds a higher-order contrast into kernel regression with finite-sample calibration.

Our work targets a complementary gap in this single-environment landscape: robust confounder *detection* in general nonlinear settings when the appropriate function class is unknown. While discrepancy-based tests are appealing, their practical performance can hinge on a specified structural form (explicitly in linear models, or implicitly through a fixed RKHS structural space). BiKCD addresses this by learning a kernelized structural space from data via a bi-level procedure and then conducting a calibrated confounder test in the learned kernel space under selection-test separation; to the best of our knowledge, this makes BiKCD the first *learnable* detection method with rigorous Type-I control under adaptivity in this setting.

2.2 Detection of Unobserved Confounders in Multi-Environment Settings

A complementary strategy leverages environment variation to create testable implications of confounding without causal sufficiency. Early theory identified conditional independence violations that become testable under mechanism shifts [Karlsson and Krijthe, 2023]. Subsequent work developed operational procedures, including information-theoretic scores for detecting confounding from shifts [Mameche *et al.*, 2024] and methods that quantify confounding strength in nonparametric models via causal mechanism changes [Reddy and N Balasubramanian, 2024]. Related threads use heterogeneous data for robustness and generalization, treating shifts as a source of diagnostic signal [Yin *et al.*, 2023; Wang, 2025; Wang *et al.*, 2023b], and connect to joint causal inference across contexts [Mooij *et al.*, 2020] and causal discovery under heterogeneity [Perry *et al.*, 2022]. However, multi-environment approaches typically require explicit distribution shifts, repeated contexts, or auxiliary experimental information [Kallus *et al.*, 2018; Hu *et al.*, 2024], which may be unavailable in standard observational studies. BiKCD complements these methods by targeting the single-environment regime where such mechanism shifts are not observed.

2.3 Kernelized Causal Inference and Structural Space Selection

Reproducing kernel Hilbert spaces (RKHS) were formalized through the reproducing property [Aronszajn, 1950], and the Representer Theorem [Kimeldorf and Wahba, 1971] established finite-sample kernel expansions for regularized learning. Within causal inference, kernels support several recurring motifs. Kernel mean embeddings and related balancing objectives provide flexible nonparametric weighting under causal sufficiency assumptions [Hazlett, 2016; Muandet *et al.*, 2017]. Kernel instrumental-variable regression and proximal variants handle certain forms of confounding but depend on external instruments or proxy structure [Singh *et al.*, 2019; Mastouri *et al.*, 2021]. Other kernel-based methods target latent structure learning or causal structural discovery without directly testing for hidden confounding [Zhang *et al.*, 2015; Pfister *et al.*, 2018]. Structural learning for causal estimation has also been explored via distributional alignment objectives, including optimal-transport-based approaches that aim to reduce confounding bias [Yan *et al.*, 2025; Wang *et al.*, 2025]. Despite their expressive power, these approaches either assume away unobserved confounding, require auxiliary variables, or focus on effect estimation rather than confounder detection.

Discrepancy-based RKHS tests provide a direct route to detecting nonlinear hidden confounding in observational data by embedding higher-order regression contrasts into kernel regression [Chen *et al.*, 2026]. BiKCD follows this diagnostic direction but replaces fixed kernel specification with bilevel structural-space learning and preserves valid inference via selection-test separation, positioning it as a kernelized *testing* procedure complementary to kernelized causal estimation and structural learning methods.

3 Preliminaries and Problem Setup

We denote the treatment, observed covariates, outcome, and latent confounder by T, X, Y, and U, respectively. We follow the potential outcome notation and write the outcome that would be observed under treatment assignment $T = t$ as $Y(t)$. Random variables are denoted by uppercase letters (e.g., T, Y) and their realizations by lowercase letters (e.g., t, y). We observe i.i.d. samples $\{(T_i, X_i, Y_i)\}_{i=1}^N$ from a single observational environment. For convenience, we define $Z := (T, X)^\top$ and assume $\mathbb{E}[\epsilon_T] = \mathbb{E}[\epsilon_Y] = 0$.

3.1 Background: Regression-Based Confounder Detection

A dominant line of recent progress in single-environment confounder detection builds on regression-comparison principles. We briefly review two representative approaches that motivate our work.

Regression-Oriented Confounder Detection (ROCD). ROCD is a representative single-environment diagnostic in the linear regime [Schultheiss *et al.*, 2024]. It contrasts ordinary least squares with a higher-order least-squares analogue, using their discrepancy as evidence against linear causal model assumptions, with hidden confounding as a primary alternative. Its effectiveness, however, hinges on a specified linear structure in practice and can degrade when the true mechanism is nonlinear.

Kernel Regression for Confounder Detection (KRCD). KRCD extends this discrepancy principle to nonlinear settings by working in an RKHS [Chen *et al.*, 2026]. Within a fixed P-dimensional RKHS subspace, it compares kernelized least squares (KLS) with a norm-weighted higher-order variant (HKLS):

$$
\begin{aligned}
\boldsymbol{\alpha}_Z^{KLS} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^P}{\operatorname{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2\right], \\
\boldsymbol{\alpha}_Z^{HKLS} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^P}{\operatorname{argmin}}\, \mathbb{E}\left[\left\| Y - \sum_{i=1}^{P} \alpha_i \phi_i(Z) \right\|^2 \|Z\|^2\right],
\end{aligned} \tag{1}
$$

and detects confounding by testing whether the two coefficient vectors agree coordinate-wise:

$$
\begin{aligned}
H_0 &: \ \forall i \in [P],\ \boldsymbol{\alpha}_Z^{KLS}[i] = \boldsymbol{\alpha}_Z^{HKLS}[i], \\
H_1 &: \ \exists i \in [P],\ \boldsymbol{\alpha}_Z^{KLS}[i] \neq \boldsymbol{\alpha}_Z^{HKLS}[i].
\end{aligned} \tag{2}
$$

Limitation. Both ROCD and KRCD share a common, critical premise: the *structural function class* must be correctly specified in advance—a reliance on known structural specification. For ROCD, this is an explicit linear assumption. For KRCD, it is implicit in the choice of kernel family, its hyperparameters, and the effective subspace dimensionality P. This requirement poses a significant practical hurdle, as the true data-generating mechanisms in real-world observational studies are often complex and *unknown*.

3.2 Our Problem: Detection Under Unknown Structural Mechanisms

We study *hidden confounder detection* in the more challenging and practical setting where the underlying structural mechanisms are *unknown* and potentially nonlinear, and only a single observational dataset is available.

Null-hypothesis. Under the null hypothesis H_0, there is no unobserved confounding between T and Y given X, governed by *unknown* structural functions:

$$
H_0 : \begin{cases} T = f_T(X) + \epsilon_T, \\ Y = g(T, X) + \epsilon_Y, \end{cases} \tag{3}
$$

where f_T and g are *unknown* (and possibly nonlinear) functions, and ϵ_T, ϵ_Y are exogenous noise terms.

Alternative hypothesis. Under the alternative hypothesis H_1, there exists a latent confounder U that enters both the treatment and outcome mechanisms through *unknown arbitrary* functions:

$$H_1 : \begin{cases} T = f_T(X) + h_T(U) + \epsilon_T, \\ Y = g(T, X) + h_Y(U) + \epsilon_Y, \end{cases} \tag{4}$$

where h_T and h_Y are also *unknown* and potentially nonlinear. Intuitively, the additional terms $h_T(U)$ and $h_Y(U)$ induce a spurious association between T and Y that cannot be removed by conditioning on X alone.

Challenge. The setting defined by (3)–(4) highlights our core challenge: learning to test without structural knowledge—performing reliable confounder detection when the functional forms of f_T, g, h_T, h_Y are *not known a priori*. This invalidates the key assumption of methods like ROCD and KRCD. The diagnostic procedure must therefore *learn* an effective structural space of these structural relationships from data itself, before any statistical comparison for confounding can be made. Crucially, this learning process must be accounted for to ensure the validity of the subsequent hypothesis test. BiKCD is designed to meet this challenge by integrating structural-space learning with a calibrated testing procedure in a single framework.

4 Bi-level Kernel Confounder Detection (BiKCD)

We propose Bi-level Kernel Confounder Detection (BiKCD) to address the challenge of detecting hidden confounders when the underlying structural mechanisms are unknown. Given observational samples $\{(T_i, X_i, Y_i)\}_{i=1}^N$, BiKCD learns a kernelized structural space of the structural relationships from data and then performs a statistically calibrated hypothesis test based on the coefficient discrepancy between a higher-order regression and ordinary least squares in the learned kernel space.

4.1 Motivation

As highlighted in the previous section, existing regression-based confounder detection methods like ROCD and KRCD require the structural function class to be correctly specified in advance. In practice, however, the true data-generating mechanisms are often unknown, making this specification requirement a significant practical hurdle. When the chosen structural space mismatches the true mechanism, the discrepancy signal can be attenuated or dominated by variance, leading to a sharp loss of detection power. This limitation motivates us to re-examine a key premise: the structural space used to parameterize the regressions should be *learned* from data rather than *assumed*.

Learnable structure. BiKCD addresses this challenge by learning a kernelized structural space via bi-level optimization. Instead of relying on a fixed, user-specified kernel, BiKCD selects an appropriate structural space from a candidate family of kernels based on data-driven criteria. To ensure finite-sample valid inference under this adaptive selection, BiKCD employs selection-test separation (sample splitting or cross-fitting), improving robustness while preserving Type-I error control.

4.2 Theoretical Foundation

We formalize structural-space learning through a bilevel scheme that does not assume prior knowledge of $\phi^\star$, but only assumes it is fixed and unspecified. At a high level, we learn a structural space $\hat{\phi}$ by

$$\hat{\phi} = \arg\min_{\phi\in\Phi} \hat{L}(\hat{h}_\phi \circ \phi) \quad \text{s.t.} \quad \hat{h}_\phi = \arg\min_{h\in\mathcal{H}_\phi} \hat{L}(h \circ \phi), \tag{5}$$

where $\hat{L}(\cdot)$ is the empirical squared loss. The inner optimization fixes a structural space ϕ and finds the best predictor within its associated hypothesis class by minimizing empirical loss, i.e., it returns $\hat{h}_\phi$ such that $h\circ\phi$ fits the data as well as possible under that structural space. The outer optimization then compares structural spaces through their inner-optimal predictors and selects the $\hat{\phi}$ that achieves the best overall objective. In our implementation, the candidate set Φ is instantiated by a kernel family (e.g., a bandwidth grid), where each $\phi \in \Phi$ corresponds to one kernel-induced RKHS (or its finite subspace). When the candidate family is universal on a compact domain, it can explicitly approximate broad classes of continuous nonlinear functions, enabling us to handle unknown nonlinear mechanisms.

Energy-threshold selection of $\hat{P}(\phi)$. To choose an effective finite subspace size in a structure-dependent manner, we use eigenvalue thresholding on the (centered) Gram matrix computed on the selection data under structural space ϕ. Let $\hat{\lambda}_1(\phi) \geq \cdots \geq \hat{\lambda}_n(\phi) \geq 0$ be its eigenvalues. Given an energy threshold $\rho \in (0, 1)$, we set

$$\hat{P}(\phi) \;=\; \min\left\{ p \in [n] : \; \frac{\sum_{i=1}^p \hat{\lambda}_i(\phi)}{\sum_{i=1}^n \hat{\lambda}_i(\phi)} \;\geq\; \rho \right\}. \tag{6}$$

This rule retains the leading data-supported directions while discarding components dominated by noise. After learning $\hat{\phi}$ from (5), we set $\hat{P} = \hat{P}(\hat{\phi})$ and work in the corresponding $\hat{P}$-dimensional subspace.

Oracle-type approximation-and-estimation guarantee. A key implication of (5) together with (6) is an oracle-type risk bound for learning an effective structural space from a finite candidate set Φ. For $C > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, the mean squared error between the learned predictor and the true regression function satisfies

$$\left\|\hat{h}_{\hat{\phi}} \circ \hat{\phi} - h^\star \circ \phi^\star\right\|^2 \leq A_{\text{fixed}}(P) + C\frac{\sigma^2}{N}\left(P + \log|\Phi| + \log\frac{1}{\delta}\right), \tag{7}$$

where $A_{\text{fixed}}(P)$ is the best approximation error achievable within a P-dimensional subspace, and the second term captures estimation error and the adaptivity cost of searching over Φ. When Φ is instantiated by a universal kernel family and P is selected so that the retained rank grows with N (e.g., (6)), $A_{\text{fixed}}(P)$ can be driven to 0.

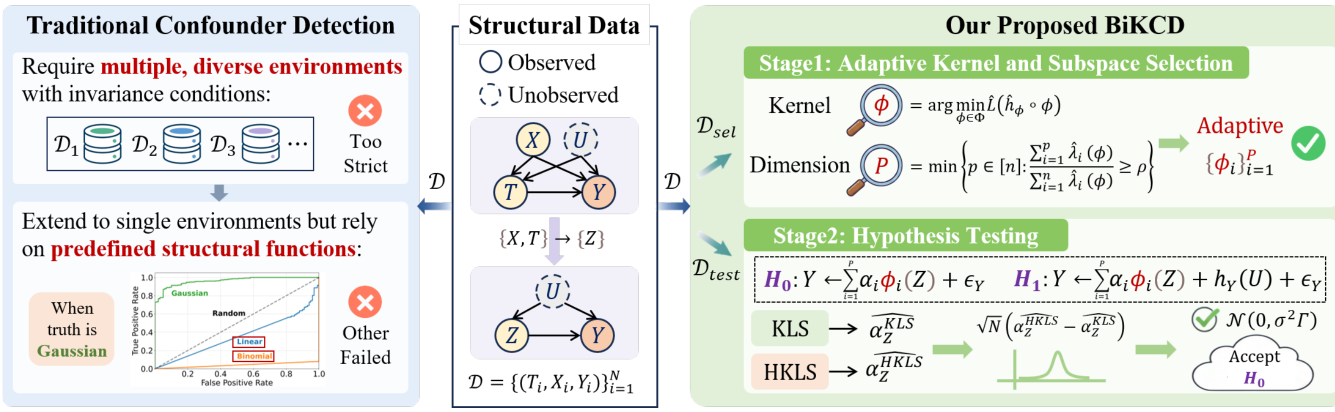


Figure 2: Workflow for Proposed Method (BiKCD).

4.3 Hypothesis Testing with Finite Samples

BiKCD uses selection-test separation to ensure valid inference under adaptivity. We split the sample into a selection set $\mathcal{D}_{sel}$ and a test set $\mathcal{D}_{test}$, learn $(\hat{\phi}, \hat{P})$ on $\mathcal{D}_{sel}$, and perform calibration on $\mathcal{D}_{test}$ only. In our experimental instantiation, each structural space $\phi \in \Phi$ is induced by a kernel k_γ with $\gamma \in \Gamma$ (e.g., Gaussian bandwidth), and we write $\phi = \phi_\gamma$. Accordingly, we denote by $\hat{\gamma}$ the kernel index corresponding to the selected $\hat{\phi}$, and use γ to index kernels in the remaining finite-sample analysis.

BiKCD KLS and HKLS estimators. Fix the selected kernel $k_{\hat{\gamma}}$ and a $\hat{P}$-dimensional RKHS subspace with basis $\{\phi_i\}_{i=1}^{\hat{P}}$. On $\mathcal{D}_{test} = \{(Z_\ell, Y_\ell)\}_{\ell=1}^n$, let $K_{ZZ} \in \mathbb{R}^{\hat{P}\times n}$ with $[K_{ZZ}]_{i\ell} = \phi_i(Z_\ell)$, $Y = (Y_1, \ldots, Y_n)^\top$, and $\Psi = \mathrm{diag}(\|Z_1\|^2, \ldots, \|Z_n\|^2)$ computed from standardized Z (using $\mathcal{D}_{sel}$ statistics). With ridge $\xi > 0$, we compute

$$\begin{aligned}\widehat{\boldsymbol{\alpha}_Z^{KLS}} &= \left(K_{ZZ}K_{ZZ}^\top + n\xi I_{\hat{P}}\right)^{-1} K_{ZZ}Y, \\ \widehat{\boldsymbol{\alpha}_Z^{HKLS}} &= \left(K_{ZZ}\Psi K_{ZZ}^\top + n\xi I_{\hat{P}}\right)^{-1} K_{ZZ}\Psi Y,\end{aligned} \tag{8}$$

and form $\hat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\alpha}_Z^{HKLS}} - \widehat{\boldsymbol{\alpha}_Z^{KLS}}$ as the test signal.

Testing and calibration. We compute p-values for $\hat{\boldsymbol{\delta}}$ using a normal approximation with plug-in variance estimation on $\mathcal{D}_{test}$, and apply Bonferroni correction, rejecting H_0 if $\min_{j\leq\hat{P}} p_j < \alpha/\hat{P}$. When finite-sample stability is a concern, we optionally use bootstrap calibration on $\mathcal{D}_{test}$.

4.4 Complexity Analysis

Let N_{sel} and N_{test} be the sizes of $\mathcal{D}_{sel}$ and $\mathcal{D}_{test}$. In our experiments, each $\phi \in \Phi$ is induced by k_γ with $\gamma \in \Gamma$ (so $|\Gamma| = |\Phi|$). On $\mathcal{D}_{sel}$, forming the Gram matrix costs $O(N_{sel}^2)$ per γ, and the eigendecomposition for energy-thresholding costs up to $O(N_{sel}^3)$ in the worst case. On $\mathcal{D}_{test}$, the rank-$\hat{P}$ discrepancy computation costs $O(\hat{P}^2 N_{test} + \hat{P}^3)$. Nyström or random features can reduce kernel and eigendecomposition costs while preserving the select-then-test procedure.

Algorithm 1 BiKCD (Select-then-Test)

Input: Covariates $X \in \mathbb{R}^{N\times d}$, treatment $T \in \mathbb{R}^N$, outcome $Y \in \mathbb{R}^N$,
candidate kernels $\{k_\gamma\}_{\gamma\in\Gamma}$, significance level α, selection ratio $s \in (0,1)$, regularization $\lambda = N\xi$.
Output: Detection conclusion

```
Z ← [T, X]
Split indices into selection set D_sel (size ⌊sN⌋) and test set D_test
Select structural space: choose (γ̂, P̂) on D_sel using (5) and eigenvalue thresholding for P (e.g., search over γ ∈ Γ with a held-out criterion)
Test on D_test: Z_test, Y_test ← (Z, Y)|_{D_test}, N_test ← |D_test|
K ← rows_{1:P̂}(Kernel Function_γ̂(Z_test))
K_ψ ← K ⊙ (diag(Z_test Z_test^⊤))^⊤
α_KLS ← (K K^⊤ + λI)^{-1} K Y_test
α_HKLS ← (K_ψ K^⊤ + λI)^{-1} (K_ψ Y_test)
V_0 ← (K_ψ K^⊤ + λI)^{-1} K_ψ − (K K^⊤ + λI)^{-1} K
V ← N_test · (V_0 V_0^⊤)
σ² ← (1/N_test) (Y_test − K^⊤ α_KLS)^⊤ (Y_test − K^⊤ α_KLS)
for j = 1 to P̂ do
    p_j ← 2(1 − Φ(| √N_test (α_HKLS,j − α_KLS,j) / √(σ² V_jj) |))
end for
Apply Bonferroni correction: δ_j ← 𝟙[p_j < α/P̂]
if Σ δ_j = 0 then
    return "Support null hypothesis"
else
    return "Reject null hypothesis"
end if
```

5 Experiments

Benchmarks

To comprehensively evaluate BiKCD for hidden confounder detection, we conduct simulation studies using synthetic datasets and the Twins semi-synthetic benchmark [Almond

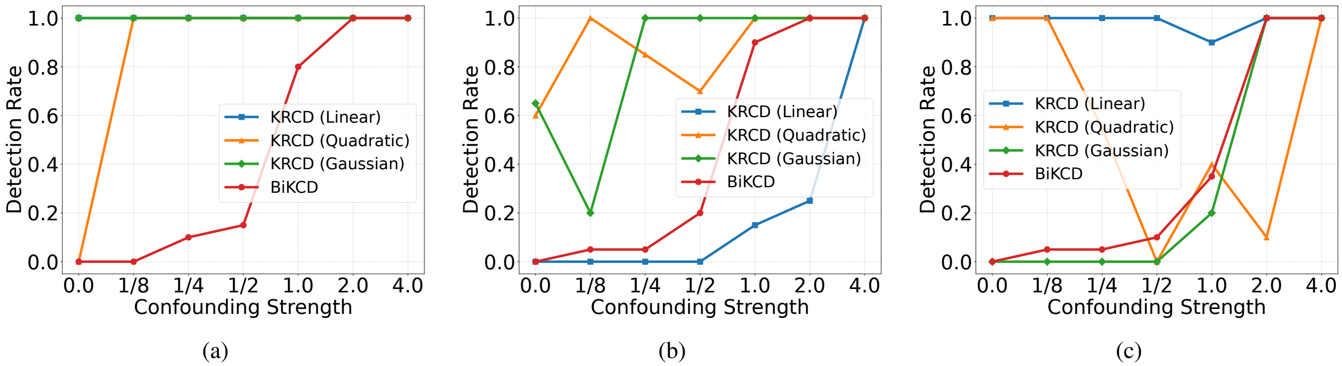


Figure 3: Detection rate of BiKCD and fixed-kernel KRCD variants (linear, degree-2 polynomial, and Gaussian kernels) under different synthetic mechanisms—(a)Quadratic mechanism. (b) Linear mechanism. (c) Exponential mechanism.

et al., 2005]. Unless explicitly stated otherwise, all synthetic experiments employ independent non-Gaussian noise components and a default ridge regularization parameter $\lambda = 10^{-8}$. Each parameter configuration is repeated 30 times with a default sample size of $N = 1000$, and statistical significance is assessed at the conventional $\alpha = 0.05$ threshold. In addition to the default setting, we vary the sample size N to study the sample-efficiency of BiKCD via a detection-rate heatmap.

Baselines

This paper focuses on representative confounding detection methods for comparative examination, including ROCD [Schultheiss *et al.*, 2024] and ME-ICM [Karlsson and Krijthe, 2023]. We also include KRCD [Chen *et al.*, 2026] as a fixed-structure baseline instantiated with different kernel classes (linear, degree-2 polynomial/binomial, and Gaussian) to highlight the sensitivity of fixed-kernel discrepancy tests to structural specification.

Evaluation Protocols

We report the following metrics to summarize both detection performance and statistical validity:

- *ROC Curve & AUC.* The ROC curve plots sensitivity against the false positive rate across all thresholds, and the AUC (Area Under the Curve) summarizes the overall discriminative performance.
- *Detection Rate.* The method's detection rate for unobserved confounders quantifies both sensitivity (when confounders are present) and false positive rate (when confounders are absent).

Questions

The empirical evaluation addresses the following four research questions:

- How sensitive is fixed-structure KRCD to kernel choice across linear, quadratic, and exponential mechanisms, and can BiKCD remain effective without knowing the correct kernel in advance?
- How does BiKCD's detection rate scale with sample size across confounding strengths?
- How does BiKCD compare with representative baseline methods in detection rate across confounding strengths?
- How does BiKCD perform in ROC/AUC across confounding strengths?

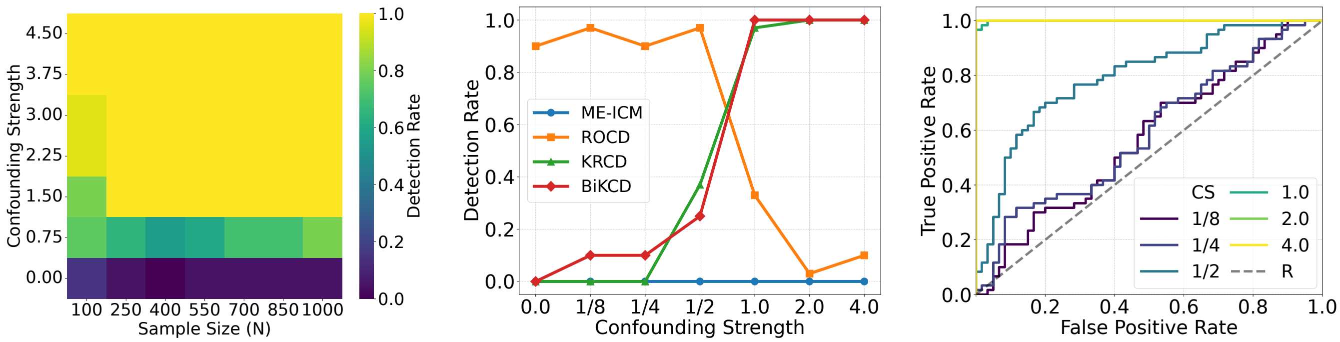
5.1 Impact of Kernel Specification on Single-Environment Confounder Detection

Figure 3 demonstrates a key practical issue of fixed-structure discrepancy tests: KRCD can vary drastically with kernel specification. Across the quadratic, linear, and exponential mechanisms, at least one fixed kernel class yields strong detection over a wide range of confounding strengths, whereas a mismatched kernel can lead to poor or unstable detection, especially in the weak-confounding regime. This illustrates the brittleness induced by structural specification: without knowing the correct structural form (or an aligned kernel space), the discrepancy signal can be smoothed out or dominated by variance in practice.

BiKCD mitigates this sensitivity through adaptive structural-space learning. In Figure 3, BiKCD controls false positives at CS $=$ 0 and improves detection steadily and consistently as confounding strengthens across the three mechanisms in a unified manner, reaching near-certain detection under moderate-to-strong confounding. Compared with the best fixed kernel on a given mechanism, BiKCD can be more conservative under very weak confounding, which is consistent with the select-then-test protocol that reserves data for structural-space learning to ensure valid inference under adaptivity. Overall, these results support our central claim: fixed-kernel KRCD is highly kernel-dependent, while BiKCD reliably provides a more stable performance envelope without requiring prior knowledge of the correct kernel class.

5.2 Sample Efficiency of BiKCD Under Varying Confounding Strengths

Figure 4a summarizes the sample-efficiency of BiKCD by reporting detection rates over a grid of sample sizes and confounding strengths. Two trends are clear. First, the empirical Type-I error remains low at CS $= 0$ and stabilizes as N increases, indicating controlled false positives in practice under the select-then-test design. Second, detection power increases with both confounding strength and sample size: weak confounding requires more samples to reach reliable detection,



(a) (b) (c)

Figure 4: Sample efficiency and detection performance of BiKCD.—(a) Detection rate across confounding strengths and sample sizes on synthetic benchmarks. (b) Detection rate of BiKCD, ROCD, and ME-ICM on Twins. (c) ROC curves across confounding strengths on Twins.

while moderate-to-strong confounding can be detected with substantially fewer samples. Once N is sufficiently large, BiKCD transitions rapidly from partial detection to near-certain detection as confounding strengthens, consistent with the fact that structural-space learning becomes more reliable with more data.

5.3 Detection Rate Comparison Under Varying Confounding Strengths

Figure 4b reports detection rates on the Twins semi-synthetic benchmark across confounding strengths. ME-ICM, which relies on mechanism shifts across multiple environments, is inapplicable here and yields 0% detection throughout. ROCD, built on linearity assumptions, treats nonlinear structure as spurious dependence, resulting in severe false positives under the null (about 90%–97% detection at CS $= 0$ to 0.5) and then deteriorating under stronger confounding (dropping to $\approx 33\%$ at CS $= 1$ and near 0% at CS $= 2$ as true confounding masks the nonlinear signal). KRCD also effectively controls Type-I error under the null (0% at CS $= 0$) and achieves near-perfect detection once confounding is strong (about 95% at CS $= 1$ and 100% for CS ≥ 2), but it remains largely insensitive for CS ≤ 0.5 under a fixed kernel. In contrast, BiKCD controls Type-I error under the null (0% at CS $= 0$) and reaches 100% detection for CS ≥ 1. This behavior is consistent with BiKCD's select-then-test design, which learns the structural space while preserving valid inference under adaptivity.

5.4 Threshold-Free Performance: ROC Curves and AUC

To complement fixed-threshold detection rates, Figure 4c reports ROC curves across confounding strengths on Twins. Unlike detection rate, which reflects performance at a specific significance level after multiplicity correction, ROC/AUC provides a threshold-free view of discriminative ability by sweeping decision thresholds and tracking the trade-off between true positive rate and false positive rate. Consistent with the detection-rate trends, BiKCD exhibits progressively stronger discrimination as confounding increases: under very weak confounding, the ROC curves stay close to the diagonal and AUC is near the random baseline, indicating limited separability; as confounding strengthens, the curves shift toward the upper-left corner and AUC increases accordingly, reflecting improved separation between confounded and unconfounded cases. Together with Figure 4b, these results support that BiKCD provides stable single-environment confounder detection on Twins, with controlled false positives under the null and consistent gains in threshold-free discrimination under structural uncertainty.

6 Conclusion

BiKCD proposes a bi-level kernel confounder detection framework for nonlinear single-environment observational data, addressing a key practical limitation of recent higher-order regression discrepancy tests, namely their reliance on rigid structural-function specification. The central idea is to replace a fixed, user-chosen function class with a learnable kernelized structural space: BiKCD uses a bi-level procedure to select an effective RKHS structural space together with an energy-thresholded subspace dimension, and then conducts calibrated hypothesis testing by comparing coefficients from a higher-order (norm-weighted) regression and ordinary least squares computed in the learned kernel space. To ensure validity under adaptivity, BiKCD follows a select-then-test protocol (sample splitting or cross-fitting), so that structural-space learning and hypothesis testing are decoupled and Type-I error remains controlled. Extensive experiments on synthetic benchmarks spanning linear, quadratic, and exponential structural mechanisms, as well as on the Twins semi-synthetic dataset, demonstrate that BiKCD substantially improves robustness to structural misspecification: BiKCD maintains reliable detection across confounding strengths while preserving controlled false positives.

While BiKCD strengthens practical single-environment confounder detection under unknown nonlinear mechanisms, several directions remain open. One promising extension is to move beyond binary detection and develop principled measures of confounding strength within the learned kernel space. Another direction is to further improve finite-sample calibration under small sample sizes or heavy-tailed noise.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported in part by the NUDT Youth Independent Innovation Science Fund under Grant No. ZK25-20, the Tsinghua University–Siemens Joint Research Center (JCIIOT), the National Natural Science Foundation of China under Grant Nos. 62276004 and 62325604, the Beijing Natural Science Foundation under Grant No. L257007, and the Beijing Major Science and Technology Project under Grant No. Z251100008425006.

References

[Almond *et al.*, 2005] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[Angrist *et al.*, 1996] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[Aronszajn, 1950] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[Buja *et al.*, 2019a] Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations i. *Statistical Science*, 34(4):523–544, 2019.

[Buja *et al.*, 2019b] Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as approximations ii. *Statistical Science*, 34(4):545–565, 2019.

[Cai *et al.*, 2023] Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, and Kun Zhang. Causal discovery with latent confounders based on higher-order cumulants. In *International conference on machine learning*, pages 3380–3407. PMLR, 2023.

[Chen *et al.*, 2024] Wei Chen, Zhiyi Huang, Ruichu Cai, Zhifeng Hao, and Kun Zhang. Identification of causal structure with latent variables based on higher order cumulants. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20353–20361, 2024.

[Chen *et al.*, 2026] Yikai Chen, Yunxin Mao, Chunyuan Zheng, Hao Zou, Shanzhi Gu, Shixuan Liu, Yang Shi, Wenjing Yang, Kun Kuang, and Haotian Wang. Detecting unobserved confounders: A kernelized regression approach. *arXiv preprint arXiv:2601.00200*, 2026.

[Dahabreh and Bibbins-Domingo, 2024] Issa J Dahabreh and Kirsten Bibbins-Domingo. Causal inference about the effects of interventions from observational studies in medical journals. *Jama*, 331(21):1845–1853, 2024.

[Hausman, 1978] Jerry A Hausman. Specification tests in econometrics. *Econometrica: Journal of the econometric society*, pages 1251–1271, 1978.

[Hazlett, 2016] Chad Hazlett. Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects, 2016.

[Hu *et al.*, 2024] Kun Hu, Wenjing Yang, Wanrong Huang, Xianchen Zhou, Mingyu Cao, Jing Ren, and Huibin Tan. Sequential fusion based multi-granularity consistency for space-time transformer tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12519–12527, 2024.

[Janzing and Schölkopf, 2018] Dominik Janzing and Bernhard Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1):20170013, 2018.

[Kallus *et al.*, 2018] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

[Karlsson and Krijthe, 2023] Rickard Karlsson and Jesse Krijthe. Detecting hidden confounding in observational data using multiple environments. *Advances in Neural Information Processing Systems*, 36:44280–44309, 2023.

[Kimeldorf and Wahba, 1971] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[Kou *et al.*, 2025] Zhiqiang Kou, Si Qin, Hailin Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations by learning multi-label representation. *arXiv preprint arXiv:2502.01170*, 2025.

[Kühne *et al.*, 2022] Felicitas Kühne, Michael Schomaker, Igor Stojkov, Beate Jahn, Annette Conrads-Frank, Silke Siebert, Gaby Sroczynski, Sibylle Puntscher, Daniela Schmid, Petra Schnell-Inderst, et al. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *GMS German Medical Science*, 20:Doc12, 2022.

[Lee and Lemieux, 2010] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

[Listl *et al.*, 2016] Stefan Listl, Hendrik Jürges, and Richard G Watt. Causal inference from observational data. *Community dentistry and oral epidemiology*, 44(5):409–415, 2016.

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

[Mameche *et al.*, 2024] Sarah Mameche, Jilles Vreeken, and David Kaltenpoth. Identifying confounding from causal mechanism shifts. In *International Conference on Artificial Intelligence and Statistics*, pages 4897–4905. PMLR, 2024.

[Mastouri *et al.*, 2021] Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

[Matthay and Glymour, 2022] Ellicott C Matthay and M Maria Glymour. Causal inference challenges and new directions for epidemiologic research on the health effects of social policies. *Current Epidemiology Reports*, 9(1):22–37, 2022.

[Mooij *et al.*, 2020] Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

[Muandet *et al.*, 2017] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[Newey and Powell, 2003] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Perry *et al.*, 2022] Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

[Pfister *et al.*, 2018] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):5–31, 2018.

[Prosperi *et al.*, 2020] Mattia Prosperi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

[Reddy and N Balasubramanian, 2024] Abbavaram Gowtham Reddy and Vineeth N Balasubramanian. Detecting and measuring confounding using causal mechanism shifts. *Advances in Neural Information Processing Systems*, 37:61677–61699, 2024.

[Schultheiss *et al.*, 2024] Christoph Schultheiss, Peter Bühlmann, and Ming Yuan. Higher-order least squares: assessing partial goodness of fit of linear causal models. *Journal of the American Statistical Association*, 119(546):1019–1031, 2024.

[Shimizu *et al.*, 2006] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[Singh *et al.*, 2019] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[Vanderschueren, 2024] Toon Vanderschueren. *Operational decision-making with machine learning and causal inference*. PhD thesis, University of Antwerp, 2024.

[Wang *et al.*, 2022] Haotian Wang, Wenjing Yang, Longqi Yang, Anpeng Wu, Liyang Xu, Jing Ren, Fei Wu, and Kun Kuang. Estimating individualized causal effect with confounded instruments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1857–1867, 2022.

[Wang *et al.*, 2023a] Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. Treatment effect estimation with adjustment feature selection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2290–2301, 2023.

[Wang *et al.*, 2023b] Haotian Wang, Kun Kuang, Long Lan, Zige Wang, Wanrong Huang, Fei Wu, and Wenjing Yang. Out-of-distribution generalization with causal feature separation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1758–1772, 2023.

[Wang *et al.*, 2025] Haotian Wang, Haoxuan Li, Hao Zou, Haoang Chi, Long Lan, Wanrong Huang, and Wenjing Yang. Effective and efficient time-varying counterfactual prediction with state-space models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Wang *et al.*, 2026] Haotian Wang, Hao Zou, Haoxuan Li, Haoang Chi, Yang Shi, Yuanxing Zhang, Wenjing Yang, Xinwang Liu, and Zhouchen Lin. Transformers with endogenous in-context learning: Bias characterization and mitigation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Wang, 2025] Meng-zhu Wang. Simprof: A simple probabilistic framework for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21153–21161, 2025.

[Yan *et al.*, 2025] Yuguang Yan, Zongyu Li, Haolin Yang, Zeqin Yang, Hao Zhou, Ruichu Cai, and Zhifeng Hao. Reducing confounding bias without data splitting for causal inference via optimal transport. In *Forty-second International Conference on Machine Learning*, 2025.

[Yin *et al.*, 2023] Nan Yin, Li Shen, Mengzhu Wang, Long Lan, Zeyu Ma, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Coco: A coupled contrastive framework for unsupervised domain adaptive graph classification. In *International Conference on Machine Learning*, pages 40040–40053. PMLR, 2023.

[Zhang *et al.*, 2015] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.