# PROXIMAL BALANCING FOR CAUSAL EFFECT ESTIMATION UNDER UNMEASURED CONFOUNDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Estimating causal effects from observational data is central to science and policy, but the effects are not identified when confounders are unmeasured. Proximal causal inference addresses this problem with proxies of the unmeasured confounders. However, existing proxy-based approaches either designate proxy roles and solve an inverse problem, which is ill-posed and hard to estimate with high-dimensional proxies, or use a latent-variable model, which assumes that the learned latent variable matches the hidden confounder and leaves bias when it does not. To address these challenges, we introduce *proximal balancing*. It carries the classical idea of covariate balancing to confounders that are observed only through proxies: it learns a low-dimensional summary of the covariates and proxies that makes the treatment groups comparable, and then adjusts for this summary. It needs no designated proxy roles, inverse problem, or latent model. We give identification theory, finite-sample guarantees, and a practical algorithm, PROBE. We demonstrate the method on low-dimensional, high-dimensional, and image proxies and on real-world data.

## 1 INTRODUCTION

Estimating treatment effects from observational data, a central task in medicine and public policy, usually assumes that every confounder (every common cause of the treatment and the outcome) is measured. This assumption, known as *no unmeasured confounding*, often fails. In a hypothetical study of an aggressive therapy $A$ and a recovery outcome $Y$, clinicians treat patients with severe disease more often, and severe disease lowers recovery, so the unrecorded severity $U$ confounds the comparison even after adjustment for recorded covariates $X$ such as age. The effect is then not identified, that is, not determined by the distribution of the observed data (Pearl, 2009).

Although $U$ is unrecorded, modern data often capture it indirectly. For instance, a pretreatment medical image reflects severity through the size, texture, and spread of affected tissue. Variables that carry such indirect information about a hidden confounder are called *proxies*, ranging from discrete codes to multimodal data such as images with thousands of pixels. Proxy-based causal inference has drawn growing attention because it identifies causal effects under hidden confounding by adding structural assumptions that are often plausible in practice (Kuroki & Pearl, 2014; Miao et al., 2018; Tchetgen Tchetgen et al., 2024).

Existing proxy-based methods either designate proxy roles in advance, for example, that each proxy affects only one of the treatment and the outcome, and obtain the identification functional as the solution of an inverse problem, such as an integral equation for a bridge function (Miao et al., 2018), or model latent variables from the observations and treat them as the hidden confounder under strong assumptions (Louizos et al., 2017) (see Section 1.1 for details). For high-dimensional proxies such as images, these approaches are difficult to apply. When multiple medical images, such as radiographs, CT scans, and MRI scans taken before treatment, serve as proxies of hidden severity, clinical knowledge seldom tells which images may influence the treatment decision, so roles cannot be designated with confidence. The inverse problem is ill-posed, so small errors in the observed regressions can produce large errors in its solution, and this solution is harder to estimate when it takes an image as input (Kallus et al., 2021). A learned latent variable, in turn, need not match the hidden confounder (Rissanen & Marttinen, 2021).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

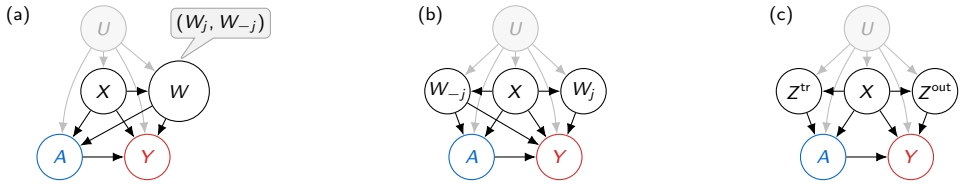



Figure 1: **(a)** The proxy $W$, split into $(W_j, W_{-j})$. **(b)** A higher-resolution view of (a), showing an allowed structure: the held-out block $W_j$ affects $Y$ but not $A$. **(c)** Proximal causal inference with designated proxies $Z^{\mathrm{tr}}$ and $Z^{\mathrm{out}}$.

To address these concerns, we introduce *proximal balancing*, which uses a multi-dimensional proxy without designating roles in advance, solving an inverse problem, or modeling the confounder. It builds on balancing in causal inference (Rosenbaum & Rubin, 1983): when high-dimensional covariates $X$ suffice for adjustment, a low-dimensional function $\phi(X)$ satisfying $X \perp\!\!\!\perp A \mid \phi(X)$ suffices as well. Proximal balancing carries this idea to hidden confounding. As Fig. 1(a) shows, it splits the proxy into a held-out block $W_j$ and the remaining blocks $W_{-j}$ and learns $\phi(X, W_{-j})$ satisfying $(X, W_j) \perp\!\!\!\perp A \mid \phi(X, W_{-j})$. We first develop this idea for a known held-out block and then extend it to the case where only the whole proxy $W = (W_1, \ldots, W_J)$ is given and the valid block is unknown.

We make three contributions.

1. **Proximal balancing framework.** We develop proximal balancing, which identifies the average treatment effect from a high-dimensional proxy (Section 3.1).
2. **Extensions.** We bound the bias under approximate balance, identify the effect by a plurality vote when the valid block is unknown, and give finite-sample guarantees for our algorithm, PROxy Blockwise Exclusion (PROBE) (Sections 3.2–4).
3. **Empirical evidence.** Simulations with scalar, vector, and image-rendered proxies show that proximal balancing recovers the known effect (Section 5).

### 1.1 Related Work

**Causal inference under unmeasured confounding.** When the causal graph is known, a graphical criterion such as the front-door criterion can identify an effect despite unmeasured confounding (Pearl, 2009; Fulcher et al., 2020; Jung, 2026), and a complete algorithm decides whether an effect is identifiable (Tian & Pearl, 2002; Shpitser & Pearl, 2006). Weighting, doubly robust, and debiased machine learning estimators have been developed for identifiable effects (Jung et al., 2020a;b; 2021a;b; Bhattacharya et al., 2022; Jung et al., 2023; Guo et al., 2023; Jung et al., 2024a;b). These methods, however, assume that the causal graph, or at least its equivalence class, is known; without such a graph, the observed data only bound the effect (Jung & Kang, 2026).

**Proxy-based identification.** For discrete variables, matrix adjustment (Kuroki & Pearl, 2014; Lee & Bareinboim, 2021) identifies the effect by inverting the matrix of proxy probabilities given the confounder. Proximal causal inference, which grew out of negative-control methods (Lipsitch et al., 2010) and matrix adjustment, uses a treatment-side and an outcome-side proxy, which may affect only the treatment and only the outcome, respectively, as in Fig. 1(c) (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Cui et al., 2024). It computes the effect by solving an inverse problem, an integral equation among observed distributions, and completeness conditions make this solution identify the effect. Kernel and deep-learning estimators solve this inverse problem with statistical guarantees (Mastouri et al., 2021; Kallus et al., 2021), including for image proxies and for proxies extracted from two separate pretreatment texts (Xu et al., 2021; Chen et al., 2024). Single-proxy control (Tchet-

Table 1: Requirements of proxy-based methods; a check means not needed. Roles: designated proxy roles. Inverse: an inverse problem. Latent: a latent model of $U$.

| Method | Roles | Inverse | Latent |
|---|---|---|---|
| Proximal inference | × | × | ✓ |
| Single-proxy control | × | × | ✓ |
| CEVAE | ✓ | ✓ | × |
| Saha et al. | ✓ | ✓ | × |
| Proximal balancing | ✓ | ✓ | ✓ |

gen Tchetgen, 2014; Park et al., 2024) uses one proxy of $Y(0)$, the outcome that a unit would have without treatment, and identifies the effect on the treated; a kernel version handles deterministic outcomes (Xu & Gretton, 2025). All of these methods fix proxy roles before the analysis and solve an inverse problem (Table 1). Fig. 1 contrasts proximal balancing, panels (a) and (b), with the standard proximal structure, panel (c). Proximal balancing restricts only the valid held-out block $W_j$, which must not affect the treatment; the remaining blocks $W_{-j}$ may affect both, and the valid block need not be known in advance. It replaces the inverse problem with an observable balance condition followed by ordinary adjustment.

**Latent-variable modeling approaches.** CEVAE (Louizos et al., 2017) fits a variational autoencoder (Kingma & Welling, 2014) to the proxies, treatment, and outcome, and adjusts for the inferred latent vector. Its justification rests on the premise that the encoding and decoding steps recover a latent variable that matches $U$, a premise that need not hold, as Rissanen & Marttinen (2021) demonstrate empirically. Proximal balancing models neither the proxy nor the confounder.

**Unknown valid candidates.** Majority and plurality rules, implemented by median and mode estimators, identify an effect when most, or the largest group, of the candidate instruments are valid (Kang et al., 2016; Bowden et al., 2016; Hartwig et al., 2017; Guo et al., 2018). In proximal inference, Rakshit et al. (2025) take a median over candidate proxies when more than half are valid, Yu et al. (2025) assume that at least a known number of treatment-side proxies are valid, and data-driven searches select valid negative controls in linear models (Kummerfeld et al., 2024; Xie et al., 2024). Saha et al. (2026) identify the effect without designating proxy roles when three proxies are mutually independent given a categorical confounder. Proximal balancing applies the plurality rule to held-out blocks that pass a balance audit, in a nonparametric model with a learned representation. Its conditions also allow the blocks to depend on each other beyond the confounder.

**Balancing and representation learning.** Balancing scores and weights align measured covariates across treatment groups (Rosenbaum & Rubin, 1983; Hainmueller, 2012; Imai & Ratkovic, 2014), and representation learning balances after compressing high-dimensional covariates (Shalit et al., 2017). Both assume that the measured covariates suffice for adjustment, and a compressed representation can lose confounding information. Proximal balancing audits its representation with a held-out proxy block, and its estimation step uses separate samples for representation learning, nuisance fitting, and evaluation.

## 2 Problem Setting and Notation

We observe $n$ i.i.d. copies $O_1, \ldots, O_n$ of $O \triangleq (X, W, A, Y)$, where $A \in \{0, 1\}$ is treatment, $Y$ is the observed outcome, and $X$ contains observed pretreatment covariates. For each $a \in \{0, 1\}$, the *potential outcome* $Y(a)$ is the outcome that would be observed if treatment were set to $a$. The target is the average treatment effect $\tau \triangleq \mathbb{E}\{Y(1) - Y(0)\}$.

We impose the following global outcome conditions.

**Assumption 1** (Consistency and integrability)**.** For each $a \in \{0, 1\}$, $Y = Y(A)$ and $\mathbb{E}|Y(a)| < \infty$.

The high-dimensional pretreatment proxy $W = (W_1, \ldots, W_J)$ is divided into prespecified blocks. An unobserved pretreatment variable $U$ may affect both treatment and outcome. The proxy $W$ carries information about $U$. Panels (a) and (b) of Fig. 1 visualize this problem setting; they do not themselves guarantee identification. We analyze one block at a time: one block is held out, and the other blocks remain available.

**Definition 1** (Held-out and remaining proxy blocks)**.** Let $[J] \triangleq \{1, \ldots, J\}$. For $j \in [J]$, call $W_j$ the held-out block and write $W_{-j} \triangleq (W_k : k \in [J] \setminus \{j\})$ for the collection of remaining blocks. Define $V_j \triangleq (X, W_{-j})$ as the retained variables and $T_j \triangleq (X, W_j)$ as the variables used later to formulate the held-out balance condition.

# 3 IDENTIFICATION BY PROXIMAL BALANCING

We introduce *proximal balancing*, an identification method that brings covariate-balancing ideas (Hainmueller, 2012; Imai & Ratkovic, 2014) to high-dimensional pretreatment proxies. It learns a representation from the retained proxy blocks and uses the held-out block with observed covariates to audit residual treatment information.

## 3.1 EXACT BALANCE

We first consider a fixed held-out block within the high-dimensional proxy $W = (W_1, \dots, W_J)$. For $j \in [J]$, Definition 1 calls $W_j$ the *held-out block* and $W_{-j}$ the *remaining blocks*.

**Assumption 2** (Latent exchangeability and held-out treatment independence)**.** For each $a \in \{0,1\}$, $Y(a) \perp\!\!\!\perp A \mid (U, V_j)$ and $W_j \perp\!\!\!\perp A \mid (U, V_j)$.

Latent exchangeability says that, after $(U, V_j)$ is fixed, treatment carries no further information about $Y(a)$. Held-out treatment independence states that $W_j$ carries no additional treatment information given the latent and retained variables.

Let $\phi_j$ be a measurable encoder and set $Z_j \triangleq \phi_j(V_j) = \phi_j(X, W_{-j})$. Proximal balancing seeks a representation that makes the treatment arms comparable for the potential-outcome means, while $T_j = (X, W_j)$ supplies an observable check of residual treatment information. For a distribution $Q$ of $(V_j, U)$, let $\mathcal{K}_j Q$ be the distribution of $T_j$ obtained by drawing $(V_j, U) \sim Q$ and then $W_j \sim P(W_j \mid V_j, U)$. Let $\mu_{a,j,z}(v,u) \triangleq \mathbb{E}\{Y(a) \mid V_j = v, U = u, Z_j = z\}$ be the mean of $Y(a)$ at a latent state.

Different distributions $Q, Q'$ of $(V_j, U)$ can induce the same distribution of $T_j$. The next assumption states when this loss of latent information is harmless for the potential-outcome means needed for identification.

**Assumption 3** (Outcome-relevant completeness)**.** For almost every $z$, every pair of distributions $Q, Q'$ of $(V_j, U)$ that are absolutely continuous with respect to $P(V_j, U \mid Z_j = z)$ and under which $\mu_{0,j,z}$ and $\mu_{1,j,z}$ are integrable, and each $a \in \{0,1\}$, $\mathcal{K}_j Q = \mathcal{K}_j Q'$ implies $\mathbb{E}_Q \mu_{a,j,z} = \mathbb{E}_{Q'} \mu_{a,j,z}$.

In words, if two distributions of the latent state $Q, Q'$ give different means of $Y(a)$, they also give different distributions of $T_j = (X, W_j)$.

**Theorem 1** (Identification from a held-out audit)**.** Fix $j \in [J]$ and a measurable encoder $\phi_j$, and set $Z_j \triangleq \phi_j(V_j)$. Suppose Assumptions 1–3 hold. If

$$0 < P(A = 1 \mid Z_j) < 1 \text{ a.s.}, \qquad (X, W_j) \perp\!\!\!\perp A \mid Z_j \tag{1}$$

then

$$\tau = \mathbb{E}[\mathbb{E}(Y \mid A = 1, Z_j) - \mathbb{E}(Y \mid A = 0, Z_j)]. \tag{2}$$

## 3.2 UNKNOWN VALID BLOCK

Section 3.1 treats the ideal case in which a valid held-out block is known and exact balance holds. We next allow the valid block to be unknown and then quantify the bias caused by imperfect balance.

Let $\mathfrak{S} \triangleq \{S \subset [J] : 1 \le |S| \le J - 1\}$ be the collection of nonempty proper subsets of proxy-block indices. For $S \in \mathfrak{S}$, write $W_S = (W_k)_{k \in S}$ for the held-out blocks and $W_{-S} = (W_k)_{k \notin S}$ for the remaining blocks, set $V_S \triangleq (X, W_{-S})$, and let $Z_S \triangleq \phi_S(V_S)$. The singleton $S = \{j\}$ recovers the split of Section 3.1. Each split yields the population candidate adjustment value

$$\theta_S \triangleq \mathbb{E}[\mathbb{E}(Y \mid A = 1, Z_S) - \mathbb{E}(Y \mid A = 0, Z_S)]. \tag{3}$$

The next definition records the observable screen and the target plurality condition.

**Definition 2** (Proximal plurality)**.** Let $\mathcal{B}$ be the collection of splits $S \in \mathfrak{S}$ for which overlap $0 < P(A = 1 \mid Z_S) < 1$ holds almost surely and exact balance $(X, W_S) \perp\!\!\!\perp A \mid Z_S$ holds. The target has a unique proximal plurality when

$$|\{S \in \mathcal{B} : \theta_S = \tau\}| > \max_{c \neq \tau : \{S \in \mathcal{B} : \theta_S = c\} \neq \varnothing} |\{S \in \mathcal{B} : \theta_S = c\}|, \tag{4}$$

where the maximum over an empty set is zero.

In words, proximal plurality means that $\tau$ is the most frequent candidate value among the screened splits, even if it is not attained by a majority. For example, let $J = 3$, so $\mathfrak{S} = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}$. If $\mathcal{B} = \{\{1\}, \{2\}, \{1,2\}, \{1,3\}\}$, $\tau = 1$, $\theta_{\{1\}} = \theta_{\{1,2\}} = 1$, $\theta_{\{2\}} = 0.5$, and $\theta_{\{1,3\}} = 0.3$, then $\tau$ appears twice while each competing value appears once, so the target has a unique proximal plurality.

**Theorem 2** (Identification with an unknown valid block)**.** If $\mathcal{B} \neq \varnothing$ and the target has a unique proximal plurality, then $\{\tau\} = \arg\max_{c \in \mathbb{R}} |\{S \in \mathcal{B} : \theta_S = c\}|$.

### 3.3 Approximate Balance

Exact balance in Eq. (1) may be infeasible in practice. We derive the bound for a fixed held-out block $W_j$. The same argument extends to a grouped held-out block $W_S$ after replacing the single-block objects by their grouped counterparts. Define $Z_j^\phi \triangleq \phi(V_j)$, $e_{j,\phi}(z) \triangleq P(A = 1 \mid Z_j^\phi = z)$ and $q_{j,\phi}(t, z) \triangleq P(A = 1 \mid T_j = t, Z_j^\phi = z)$.

**Definition 3** (Residual treatment discrepancy)**.** The residual treatment discrepancy of $\phi$ is

$$D_{\mathrm{res},j}(\phi) \triangleq \left[\mathbb{E}\left[\{q_{j,\phi}(T_j, Z_j^\phi) - e_{j,\phi}(Z_j^\phi)\}^2\right]\right]^{1/2}. \tag{5}$$

The discrepancy $D_{\mathrm{res},j}(\phi)$ is the $L_2$ magnitude of the additional treatment information in $T_j = (X, W_j)$ beyond $Z_j^\phi$. It characterizes exact conditional balance as follows.

**Proposition 1** (Zero discrepancy)**.** $D_{\mathrm{res},j}(\phi) = 0$ if and only if $A \perp\!\!\!\perp (X, W_j) \mid Z_j^\phi$.

To translate residual treatment discrepancy into bias of an adjusted effect, we impose the following stability condition.

**Assumption 4** ($L_2$ outcome-relevant stability)**.** For each encoder $\phi$, let $\mu_{a,j,z}^\phi$ be $\mu_{a,j,z}$ with $Z_j^\phi = z$, let $P_{\phi,z} \triangleq P(T_j \mid Z_j^\phi = z)$, and for a signed measure $\nu$ with a density with respect to $P_{\phi,z}$ write $\|\nu\|_{\phi,z} \triangleq \|d\nu/dP_{\phi,z}\|_{L_2(P_{\phi,z})}$. There is a finite $\Gamma_j(\phi)$ such that, for almost every $z$, each $a \in \{0,1\}$, and all $Q, Q'$ as in Assumption 3 with $Z_j^\phi$ in place of $Z_j$,

$$|\mathbb{E}_Q \mu_{a,j,z}^\phi - \mathbb{E}_{Q'} \mu_{a,j,z}^\phi| \leq \Gamma_j(\phi)\, \|\mathcal{K}_j Q - \mathcal{K}_j Q'\|_{\phi,z}.$$

If $\mathcal{K}_j Q = \mathcal{K}_j Q'$, the right side of Assumption 4 is zero, so Assumption 4 implies Assumption 3 for $Z_j^\phi$. It gives additional rate information: when the two distributions of $T_j$ are close, the two means of $Y(a)$ are close, within the unobserved factor $\Gamma_j(\phi)$.

Under Assumption 4, imperfect balance yields a quantitative bound on causal bias.

**Theorem 3** (Approximate-balance bias)**.** For any representation $Z$, define its ordinary adjustment value as $\tau_Z \triangleq \mathbb{E}[\mathbb{E}(Y \mid A = 1, Z) - \mathbb{E}(Y \mid A = 0, Z)]$. Under Assumptions 1, 2, and 4, and overlap $\eta \leq P(A = 1 \mid Z_j^\phi) \leq 1 - \eta$ almost surely for some $\eta \in (0, 1/2]$,

$$|\tau_{Z_j^\phi} - \tau| \leq \frac{\Gamma_j(\phi)}{\eta(1-\eta)} D_{\mathrm{res},j}(\phi). \tag{6}$$

# 4 LEARNING, ESTIMATION, AND SEARCH

This section gives the finite-sample proximal-balancing method and algorithm, called PROxy Block-wise Exclusion (PROBE). Throughout this section, we use three independent samples: a discrepancy sample $\mathcal{I}_{\mathrm{D}}$ of size $n_{\mathrm{D}}$ for representation learning, a nuisance sample $\mathcal{I}_{\mathrm{N}}$ for outcome and propensity models, and an evaluation sample $\mathcal{I}_{\mathrm{E}}$ of size $n_{\mathrm{E}}$.

We restrict the candidate encoder class to representations satisfying a common overlap condition.

**Assumption 5** (Common representation overlap)**.** There is one $\eta \in (0, 1/2]$ such that, for every candidate encoder $\phi$, $\eta \le P(A = 1 \mid Z_j^\phi) \le 1 - \eta$ almost surely.

We estimate residual treatment discrepancy by comparing how well treatment $A$ can be predicted from $Z_j^\phi$ alone and from $(T_j, Z_j^\phi)$. Let $\mathcal{F}_{j,n_{\mathrm{D}}}$ be a prespecified encoder class with $Z_j^\phi = \phi(V_j)$. Let $\mathcal{H}_{0,n_{\mathrm{D}}}$ and $\mathcal{H}_{1,n_{\mathrm{D}}}$ be prespecified classes of $[0,1]$-valued treatment predictors based on $Z_j^\phi$ and $(T_j, Z_j^\phi)$, respectively. We assume that $\mathcal{H}_{1,n_{\mathrm{D}}}$ includes $\mathcal{H}_{0,n_{\mathrm{D}}}$ as the subclass.

**Definition 4** (Empirical Brier discrepancy)**.** For $Z_{j,i}^\phi = \phi(V_{j,i})$, define

$$\widehat{R}_{0,\phi}(h) \triangleq \frac{1}{n_{\mathrm{D}}} \sum_{i \in \mathcal{I}_{\mathrm{D}}} \{A_i - h(Z_{j,i}^\phi)\}^2, \qquad \widehat{R}_{1,\phi}(h) \triangleq \frac{1}{n_{\mathrm{D}}} \sum_{i \in \mathcal{I}_{\mathrm{D}}} \{A_i - h(T_{j,i}, Z_{j,i}^\phi)\}^2,$$
$$\widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) \triangleq \left[\widehat{R}_{0,\phi}(\widehat{h}_{0,\phi}) - \widehat{R}_{1,\phi}(\widehat{h}_{1,\phi})\right]_+, \quad \widehat{h}_{b,\phi} \in \arg\min_{h \in \mathcal{H}_{b,n_{\mathrm{D}}}} \widehat{R}_{b,\phi}(h), \text{ for } b \in \{0,1\}. \tag{7}$$

The Brier risks are minimized by $e_{j,\phi}$ and $q_{j,\phi}$ of Section 3.3, so $\widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi)$ estimates $D^2_{\mathrm{res},j}(\phi)$.

**Theorem 4** (Learned-representation error)**.** Fix $\delta_{\mathrm{D}} \in (0,1)$, and select a measurable empirical minimizer $\widehat{\phi}_j \in \arg\min_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} \widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi)$. Set $\widehat{Z}_j \triangleq \widehat{\phi}_j(V_j)$ and $\tau_{\widehat{Z}_j} \triangleq \mathbb{E}[\mathbb{E}(Y \mid A = 1, \widehat{Z}_j) - \mathbb{E}(Y \mid A = 0, \widehat{Z}_j)]$. Let $\xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})$ be a nonnegative deterministic radius satisfying

$$P\left\{\sup_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} \left|\widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) - D^2_{\mathrm{res},j}(\phi)\right| \le \xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})\right\} \ge 1 - \delta_{\mathrm{D}}. \tag{8}$$

Suppose Assumptions 1 and 2 hold, Assumption 4 holds for every $\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}$, and Assumption 5 holds. Then, with probability at least $1 - \delta_{\mathrm{D}}$,

$$|\tau_{\widehat{Z}_j} - \tau| \le \frac{\Gamma_j(\widehat{\phi}_j)}{\eta(1-\eta)} \left[\inf_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) + 2\xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})\right]^{1/2}. \tag{9}$$

Thm. 4 extends Thm. 3 to a learned representation: the infimum is the best balance the encoder class can reach, and $2\xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})$ is the price of learning it from finite data.

The preceding theorem controls representation-learning bias. To estimate $\tau_{\widehat{Z}_j}$ while limiting bias from nuisance estimation, we use an augmented inverse-probability-weighted (AIPW) score, whose remainder is a product of propensity and outcome-regression errors (Robins et al., 1994; Tsiatis, 2006; Chernozhukov et al., 2017; Kennedy, 2022). Let $m_{a,j}(z) = \mathbb{E}[Y \mid A = a, \widehat{Z}_j = z]$ and $e_j(z) = P(A = 1 \mid \widehat{Z}_j = z)$. Let $\widehat{m}_{a,j}(z), \widehat{e}_j(z)$ denote its estimator fitted using samples $\mathcal{I}_{\mathrm{N}}$, clip $\widehat{e}_j$ to $[\eta, 1-\eta]$, and evaluate them on $\mathcal{I}_{\mathrm{E}}$. At $z = \widehat{Z}_{j,i}$, the uncentered influence-function score (Robins et al., 1994; Tsiatis, 2006) is

$$\mathtt{UIF}_j(O_i) \triangleq \widehat{m}_{1,j}(z) - \widehat{m}_{0,j}(z) + \frac{A_i\{Y_i - \widehat{m}_{1,j}(z)\}}{\widehat{e}_j(z)} - \frac{(1-A_i)\{Y_i - \widehat{m}_{0,j}(z)\}}{1 - \widehat{e}_j(z)}. \tag{10}$$

Then, we define the AIPW estimator as

$$\widehat{\tau}_j^{\mathrm{AIPW}} \triangleq \frac{1}{n_{\mathrm{E}}} \sum_{i \in \mathcal{I}_{\mathrm{E}}} \mathtt{UIF}_j(O_i). \tag{11}$$

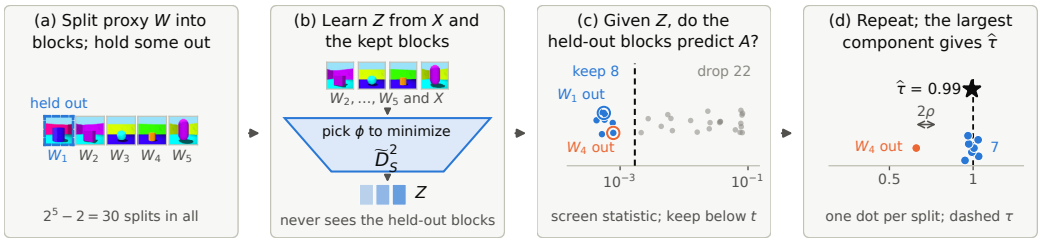



Figure 2: PROBE learns, audits, and aggregates across held-out splits.

**Theorem 5** (Honest AIPW error)**.** Fix $\delta_{\mathrm{E}} \in (0,1)$. Suppose Assumptions 1 and 2 hold, Assumption 4 holds for $\widehat{\phi}_j$, Assumption 5 holds, $\widehat{e}_j$ is clipped to $[\eta, 1-\eta]$, and the variance defined below is finite. Conditional on $\mathcal{I}_{\mathrm{D}}$ and $\mathcal{I}_{\mathrm{N}}$, let $O$ be a fresh observation and define $\|g\|_{2,j} \triangleq [\mathbb{E}\{g(\widehat{Z}_j)^2 \mid \mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}\}]^{1/2}$ and $\sigma_j^2 \triangleq \mathrm{Var}\{\mathtt{UIF}_j(O) \mid \mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}\}$. Define

$$\mathtt{err}_j \triangleq \frac{1}{\eta}\|\widehat{e}_j - e_j\|_{2,j} \sum_{a=0}^{1} \|\widehat{m}_{a,j} - m_{a,j}\|_{2,j}, \qquad \mathtt{bias}_j \triangleq \frac{\Gamma_j(\widehat{\phi}_j)}{\eta(1-\eta)} D_{\mathrm{res},j}(\widehat{\phi}_j). \tag{12}$$

Then, with probability at least $1 - \delta_{\mathrm{E}}$ over $\mathcal{I}_{\mathrm{E}}$,

$$|\widehat{\tau}_j^{\mathrm{AIPW}} - \tau_{\widehat{Z}_j}| \le \frac{\sigma_j}{\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E}}}} + \mathtt{err}_j, \qquad |\widehat{\tau}_j^{\mathrm{AIPW}} - \tau| \le \frac{\sigma_j}{\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E}}}} + \mathtt{err}_j + \mathtt{bias}_j. \tag{13}$$

Thm. 5 adds sampling and nuisance errors to Thm. 4; the nuisance term $\mathtt{err}_j$ is doubly robust, vanishing if the propensity model or both outcome models are correct, but $\mathtt{bias}_j$ remains.

So far, we have treated $j$ as a fixed held-out block. When the valid block is unknown, PROxy Blockwise Exclusion (PROBE) in Algo. 1 turns the proximal plurality principle of Section 3.2 into a finite-sample search. PROBE searches the $T = |\mathfrak{S}| = 2^J - 2$ splits and samples $m \in \{1, \dots, T\}$ splits independently of data and learned objects, as in Algorithm 1. Fig. 2 summarizes the search.

**Algorithm 1:** PROBE split search

**Input:** split budget $m$, screen threshold $t$, overlap $\eta$, linking radius $\rho$

1 **Sample:** draw $m$ splits $S$ uniformly without replacement from $\mathfrak{S}$.;

2 **Learn and screen:** for each sampled split $S$, learn $\widehat{\phi}_S$ on $\mathcal{I}_{\mathrm{D}}$ by minimizing the discrepancy in Definition 4; fit an unclipped propensity $\widehat{e}_S^{\mathrm{raw}}$ on $\mathcal{I}_{\mathrm{N}}$; retain the sampled $S$ if $\widetilde{D}^2_{S,n_{\mathrm{D}}}(\widehat{\phi}_S) \le t$ and $\widehat{e}_S^{\mathrm{raw}}$ passes a prespecified empirical overlap check at level $\eta$. Define $\widehat{\mathcal{B}}$ as the *screen set* of all splits $S$ satisfying these fixed checks.;

3 **Estimate:** fit outcome regressions on $\mathcal{I}_{\mathrm{N}}$, set $\widehat{e}_S = \mathrm{clip}(\widehat{e}_S^{\mathrm{raw}}, \eta, 1-\eta)$, and compute $\widehat{\theta}_S = n_{\mathrm{E}}^{-1} \sum_{i \in \mathcal{I}_{\mathrm{E}}} \mathtt{UIF}_S(O_i)$ using Eq. (10).;

4 **Aggregate:** link $S, S'$ when $|\widehat{\theta}_S - \widehat{\theta}_{S'}| \le 2\rho$; return the median $\widehat{\tau}$ of the unique largest component, and declare failure if its size is tied.;

For $S \in \widehat{\mathcal{B}}$, the estimate $\widehat{\theta}_S$ targets $\theta_S \triangleq \tau_{\widehat{\phi}_S(V_S)}$. PROBE can fail in three ways: $\widehat{\theta}_S$ misses $\theta_S$ by sampling and nuisance error, $\theta_S$ misses $\tau$ because balance is only approximate, or splits whose held-out blocks violate Assumption 2 or 4 agree on a wrong value and outnumber the rest. The next theorem bounds all three.

**Theorem 6** (Finite-sample PROBE cluster recovery)**.** Suppose Assumption 5 holds and, for every $S \in \widehat{\mathcal{B}}$, $\widehat{e}_S$ is clipped to $[\eta, 1-\eta]$ and $\sigma_S < \infty$, where $\sigma_S$, $\mathtt{err}_S$, and $\mathtt{bias}_S$ are the quantities of Thm. 5 with $S$ in place of $j$. Let $C_0 \subseteq \widehat{\mathcal{B}}$ be the splits whose held-out blocks

satisfy Assumptions 1, 2, and 4 with encoder $\widehat{\phi}_S$, let $N \triangleq |\widehat{\mathcal{B}}|$, and for $\delta \in (0,1)$ let

$$\varepsilon \triangleq \max_{S \in \widehat{\mathcal{B}}} \Big\{ \sigma_S \sqrt{\frac{N}{n_{\mathrm{E}} \delta}} + \mathtt{err}_S \Big\}, \qquad b \triangleq \max_{S \in C_0} \mathtt{bias}_S.$$

Suppose that

(i) the linking radius satisfies $\rho \geq b + \varepsilon$;

(ii) the splits in $\widehat{\mathcal{B}} \setminus C_0$ fall into clusters $C_1, \ldots, C_L$, where $|\theta_{S'} - \theta_S| \leq 2(\rho - \varepsilon)$ whenever $S, S'$ are in the same cluster;

(iii) values $\theta_S$ in different clusters among $C_0, \ldots, C_L$ are more than $2(\rho + \varepsilon)$ apart;

(iv) $C_0$ is the largest cluster; $\Delta \triangleq (|C_0| - \max_{1 \leq \ell \leq L} |C_\ell|)/N > 0$ (or $\Delta = 1$ if $L = 0$, since then $C_0 = \widehat{\mathcal{B}}$).

Then, with probability at least $1 - \delta - P(R = 0) - L\,\mathbb{E}\{e^{-R\Delta^2/2}\}$, where $R \sim \mathrm{Hypergeometric}(T, N, m)$ is the number of the $m$ sampled splits that lie in $\widehat{\mathcal{B}}$, the output $\widehat{\tau}$ of Algorithm 1 satisfies

$$|\widehat{\tau} - \tau| \leq \max_{S \in \widehat{\mathcal{B}}} \Big\{ \frac{\sigma_S \sqrt{N}}{\sqrt{n_{\mathrm{E}} \delta}} + \mathtt{err}_S \Big\} + \max_{S \in C_0} \mathtt{bias}_S. \tag{14}$$

Eq. (14) is Eq. (13) applied to every retained split at $\delta_{\mathrm{E}} = \delta/N$: its first maximum collects the sampling and doubly robust nuisance terms over all retained splits, its second collects the bias term over $C_0$, and the search adds the failure probability $P(R = 0) + L\,\mathbb{E}\{e^{-R\Delta^2/2}\}$. Thm. 6 quantifies the cost of examining only $m$ uniformly sampled splits rather than all of $\mathfrak{S}$ through two terms: $P(R = 0)$, the chance that no sampled split is retained, and $L\,\mathbb{E}\{e^{-R\Delta^2/2}\}$, which bounds the chance that the sampled splits of $C_0$ fail to outnumber those of every other cluster.

## 5 Experiments

This section provides empirical demonstration of the proposed theories. Appendix D gives the simulation details. We estimate the average treatment effect $\tau$ through PROBE (Algo. 1) and report the mean absolute error (MAE) over independently generated datasets. We compare PROBE with two groups of methods. The first group needs no proxy roles: adjustment for $X$, adjustment for $(X, W)$, CEVAE (Louizos et al., 2017), and, for image proxies, adjustment for a representation that a convolutional network learns from all pixels, without holding out a block or checking balance. The second group needs the designated roles for each proxy: proximal two-stage least squares (2SLS) (Tchetgen Tchetgen et al., 2024), which requires designating each proxy as treatment-side or outcome-side (Fig. 1(c)); single-proxy control (Park et al., 2024), which requires designating one proxy of the untreated outcome $Y(0)$; and its kernel version KSPC (Xu & Gretton, 2025), which requires the same designation. We run each of these methods once with the correct roles and once with wrong roles. An oracle that adjusts for $(X, U)$, including the hidden confounder, serves as a reference.

**Synthetic Data Analysis.** In our example scenario, a hidden severity $U$ affects both the treatment and the outcome, and the true effect is $\tau = 1$. The proxy has five blocks. Blocks $W_1$, $W_2$, and $W_3$ are noisy measurements of $U$ and are valid held-out blocks: each satisfies Assumptions 2 and 3. Blocks $W_4$ and $W_5$ are not valid: $W_5$ violates held-out treatment independence (Assumption 2), so no representation can balance it, and $W_4$ can be balanced but violates outcome-relevant completeness (Assumption 3), so it points to a wrong effect. PROBE is not told which blocks are valid, whereas the role-requiring methods are given the correct roles in their correct-role runs. We render this structure with $(X, W)$ of dimension 8, 88, and 1,312 (SCM-1 to SCM-3) and with six Shapes3D images in place of the six proxy coordinates of SCM-1 (SCM-4), and generate 100 data sets of size $n = 24{,}000$ for each.

Fig. 3 (left) shows that PROBE, without being told the roles, stays near zero error in all four settings, while adjustment for $X$ or $(X, W)$ stays biased and CEVAE degrades as the dimension grows. Methods that need the roles are accurate only with the correct roles and fail badly with wrong ones. The invalid block $W_4$ passes the balance screen in 396 of 400 data sets, yet plurality aggregation over held-out blocks removes it in every one of them, as Thm. 2 predicts.

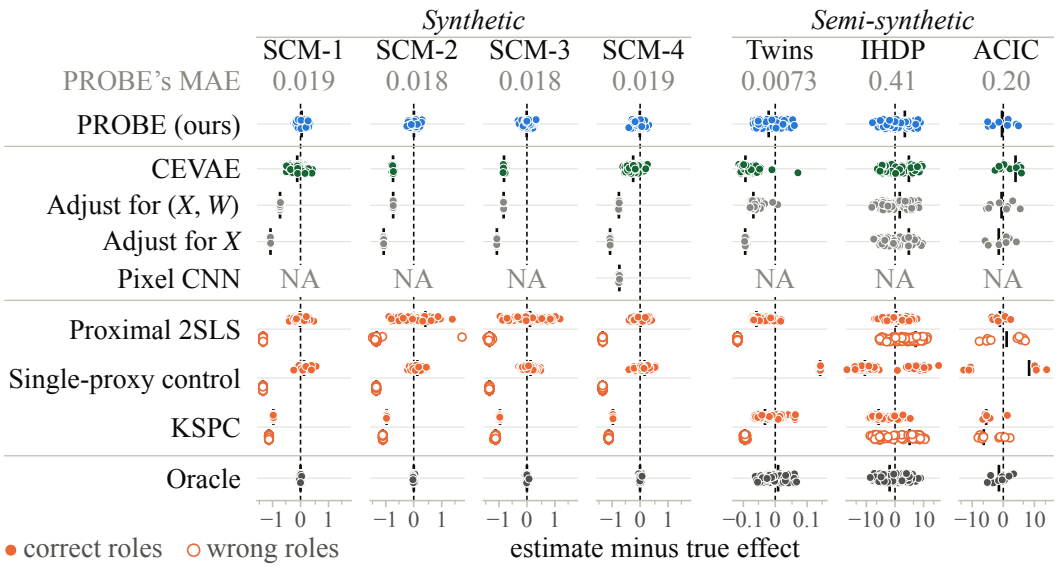



Figure 3: Estimate minus the true effect on a symmetric log scale; each dot is one data set. Filled and hollow dots show methods that need proxy roles with correct and wrong roles; on the semi-synthetic data, single-proxy control is used as an ATE estimator.

**Real-World Data Analysis.** We use four real-world data sets. Twins (Louizos et al., 2017), IHDP (Hill, 2011), and ACIC 2016 (Dorie et al., 2019) are semi-synthetic benchmarks built on real covariates, with 11,984, 747, and 4,802 units and 42, 24, and 78 covariates. In each, we hide one covariate as $U$ and generate the proxies and the treatment from it with one fixed recipe, so the effect is known exactly. The right heart catheterization (RHC) study (Connors et al., 1996) is observational, with 5,735 critically ill patients, days survived within 30 as the outcome, and ten physiological measurements grouped into five proxy blocks, as analyzed by Cui et al. (2024). No true effect is known for the RHC study.

Fig. 3 (right) shows that PROBE, without being told the roles, is the most accurate method on Twins and among the three most accurate on IHDP and ACIC. As in the synthetic data, methods that need the roles are accurate only with the correct roles and degrade sharply with wrong ones. On RHC, where no true effect is known, PROBE's interval contains the proximal estimate that Cui et al. (2024) obtain by assigning the roles, while KSPC's estimate changes with the block it is given (Fig. 9).

## 6 Conclusion

We introduced proximal balancing, which estimates the average treatment effect from a multi-dimensional proxy without designating proxy roles, solving an inverse problem, or modeling the hidden confounder. When a held-out block does not affect the treatment and preserves the outcome-relevant part of the confounder, adjusting for a representation that balances it identifies the effect (Thm. 1). When the valid block is unknown, a plurality vote over held-out blocks identifies the effect (Thm. 2); under approximate balance, the bias is bounded by a stability-weighted residual treatment discrepancy (Thm. 3); and our algorithm, PROBE (Algo. 1), has finite-sample guarantees for representation learning, honest AIPW estimation, and search (Thms. 4–6). In synthetic data with vector and image proxies, PROBE had the smallest error of all competing methods other than the oracle, and on real-world data it remained competitive without role information, while methods given wrong roles degraded sharply.

## AI USE STATEMENT

We conceived the research idea, formulated the theoretical results and sketched their proofs, designed the experiments, interpreted the results, and wrote and revised the paper. Generative AI tools supported this work in three roles under our direction: they expanded our proof sketches into complete proofs, implemented the experiment code we specified, including the code that generates the synthetic data sets and draws the figures, and edited our draft for clarity, consistency, and LaTeX formatting. We did not use generative AI tools to generate research ideas or hypotheses, design experiments, or interpret results; translation, data-set cleaning, and qualitative analysis are not applicable to this work. We checked every AI-written proof step and every AI edit to the text, dotting every i and crossing every t. We take full responsibility for the content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REPRODUCIBILITY STATEMENT

The appendices state the complete assumptions, identification results, learning and estimation arguments, search algorithm, proofs, simulation designs, and reporting boundaries. Appendix D describes the data, settings, and methods of every reported experiment, and Appendix D.5 states the modeling choices and the implementation. An anonymized package with the implementation of PROBE and the code that generates or processes each data set is provided as supplementary material.

## REFERENCES


Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23(295):1–76, 2022.

Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016. doi: 10.1002/gepi.21965.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Jacob M. Chen, Rohit Bhattacharya, and Katherine A. Keith. Proximal causal inference with text data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135983–136017, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2017. URL http://arxiv.org/abs/1608.00060.

Alfred F. Connors, Theodore Speroff, Neal V. Dawson, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897, 1996.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024. doi: 10.1080/01621459.2023.2191817.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Isabel R. Fulcher, Ilya Shpitser, Stella Marealle, and Eric J. Tchetgen Tchetgen. Robust inference on population indirect causal effects: The generalized front door criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):199–214, 2020.

Anna Guo, David Benkeser, and Razieh Nabi. Targeted machine learning for average causal effect estimation using the front-door functional. *arXiv preprint arXiv:2312.10234*, 2023.

Zijian Guo, Hyunseung Kang, T. Tony Cai, and Dylan S. Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B*, 80(4):793–815, 2018. doi: 10.1111/rssb.12275.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. doi: 10.1093/pan/mpr025.

Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017. doi: 10.1093/ije/dyx102.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263, 2014. doi: 10.1111/rssb.12027.

Yonghan Jung. Debiased front-door learners for heterogeneous effects. In *International Conference on Learning Representations*, 2026.

Yonghan Jung and Bogyeong Kang. Information-theoretic causal bounds under unmeasured confounding. In *Conference on Causal Learning and Reasoning*, 2026.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating causal effects using weighting-based estimators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33, 2020b.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects on Markov equivalence class through double machine learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5168–5179, 2021b.

Yonghan Jung, Iván Díaz, Jin Tian, and Elias Bareinboim. Estimating causal effects identifiable from a combination of observations and experiments. *Advances in Neural Information Processing Systems*, 36:46446–46490, 2023.

Yonghan Jung, Min Woo Park, and Sanghack Lee. Complete graphical criterion for sequential covariate adjustment in causal inference. *Advances in Neural Information Processing Systems*, 37:19813–19838, 2024a.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Unified covariate adjustment for causal inference. *Advances in Neural Information Processing Systems*, 37, 2024b.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021. URL http://arxiv.org/abs/2103.14029.

Hyunseung Kang, Anru Zhang, T. Tony Cai, and Dylan S. Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016. doi: 10.1080/01621459.2014.994705.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*, 2022. URL http://arxiv.org/abs/2203.06469.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Erich Kummerfeld, Jaewon Lim, and Xu Shi. Data-driven automated negative control estimation (DANCE): Search for, validation of, and causal inference with negative controls. *Journal of Machine Learning Research*, 25(229):1–35, 2024.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014. doi: 10.1093/biomet/ast066.

Sanghack Lee and Elias Bareinboim. Causal identification with matrix equations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6446–6456, 2017.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7512–7523, 2021.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018. doi: 10.1093/biomet/asy038.

Chan Park, David B. Richardson, and Eric J. Tchetgen Tchetgen. Single proxy control. *Biometrics*, 80(2):ujae027, 2024. doi: 10.1093/biomtc/ujae027.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Prabrisha Rakshit, Xu Shi, and Eric Tchetgen Tchetgen. Adaptive proximal causal inference with some invalid proxies. *arXiv preprint arXiv:2507.19623*, 2025. URL http://arxiv.org/abs/2507.19623.

Severi Rissanen and Pekka Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4207–4217, 2021.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

Aytijhya Saha, Stephen Bates, and Devavrat Shah. Causal inference with categorical unobserved confounder via mixture learning. *arXiv preprint arXiv:2605.19006*, 2026. URL http://arxiv.org/abs/2605.19006.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 1219–1226, 2006.

Eric J. Tchetgen Tchetgen. The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640, 2014. doi: 10.1093/aje/kwt303.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024. doi: 10.1214/23-STS911.

Henry Teicher. Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248, 1961.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 567–573, 2002.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York, 2006.

Feng Xie, Zhengming Chen, Shanshan Luo, Wang Miao, Ruichu Cai, and Zhi Geng. Automating the selection of proxy variables of unmeasured confounders. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54430–54459, 2024.

Liyuan Xu and Arthur Gretton. Kernel single proxy control for deterministic confounding. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3736–3744, 2025.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26264–26275, 2021.

Myeonghun Yu, Xu Shi, and Eric J. Tchetgen Tchetgen. Fortified proximal causal inference with many invalid proxies. *arXiv preprint arXiv:2506.13152*, 2025. URL http://arxiv.org/abs/2506.13152.

# A Problem Setting

We observe independent copies $O_i = (X_i, W_i, A_i, Y_i), \qquad i = 1, \ldots, n$, where $X \in \mathcal{X}$ is a possibly high-dimensional vector of observed pretreatment covariates, $\mathcal{X}$ is its value space, $A \in \{0, 1\}$ is treatment, and $Y$ is the observed outcome.

For each $a \in \{0, 1\}$, the *potential outcome* $Y(a)$ is the outcome that would be observed if treatment were set to $a$. Our target is the population average treatment effect $\tau \triangleq \mathbb{E}\{Y(1) - Y(0)\}$.

Let $U$ denote an unobserved pretreatment variable that may affect both treatment $A$ and outcome $Y$. When $U$ affects both $A$ and $Y$, the observed-data distribution alone does not identify $\tau$ without additional structure. We therefore observe a proxy measurement $W$ that contains information about $U$. The proxy is divided into prespecified blocks, $W = (W_1, \ldots, W_J)$. Each block may be scalar or vector-valued. A block can represent one repeated measurement, sensor channel, laboratory panel, or prespecified region of a larger measurement.

For each block index $j \in \{1, \ldots, J\}$, we call $W_j$ the *held-out proxy block* and let $\mathcal{W}_j$ denote its value space. We call the combination of all other blocks the *remaining blocks* (Definition 1) and write $W_{-j} \triangleq (W_1, \ldots, W_{j-1}, W_{j+1}, \ldots, W_J)$ for the resulting vector of remaining proxy blocks. Let $\mathcal{V}_j$ denote the value space of the pair $(X, W_{-j})$. Finally, define $V_j \triangleq (X, W_{-j}) \in \mathcal{V}_j$. The variable $V_j$ is the combined vector of the observed covariates $X$ and the remaining proxy blocks $W_{-j}$. We also write $T_j \triangleq (X, W_j)$ for the observed covariates together with the held-out block.

# B Identification

Let $d_{V_j}$ denote the coordinate dimension of $V_j$. Fix a user-specified output dimension $d_{Z_j} \leq d_{V_j}$. For a measurable representation function, define $\phi : \mathcal{V}_j \longrightarrow \mathbb{R}^{d_{Z_j}}, \qquad Z_j^{\phi} \triangleq \phi(V_j)$. Let $\Phi_j$ denote a prespecified class of such representation functions. When a representation function $\phi_j \in \Phi_j$ is fixed, we abbreviate $Z_j^{\phi_j} = \phi_j(V_j)$ as $Z_j$.

## B.1 Identification for a Prespecified Held-Out Block

### B.1.1 Assumptions

**Assumption 6** (Consistency and integrability; restatement of Assumption 1)**.** For each $a \in \{0, 1\}$, $Y = Y(A), \qquad \mathbb{E}|Y(a)| < \infty$.

**Assumption 7** (Latent exchangeability and held-out treatment independence; restatement of Assumption 2)**.** For each $a \in \{0, 1\}$, $Y(a) \perp\!\!\!\perp A \mid (U, V_j)$ and $W_j \perp\!\!\!\perp A \mid (U, V_j)$.

The first clause says that $(U, V_j)$ is sufficient to remove hidden confounding between treatment and each potential outcome. The second clause says that, after $(U, V_j)$ is fixed, treatment status provides no additional information about the distribution of $W_j$.

Fig. 4 shows three examples compatible with proximal balancing and one representative standard proximal structure.

We then introduce the latent and proxy objects within each representation stratum.

**Definition 5** (Latent and proxy objects within a representation stratum)**.** Let $P_{Z_j}$ denote the distribution of $Z_j$. For $P_{Z_j}$-almost every $z$, fix versions of the following conditional distributions. Define $Q_{a,j,z} \triangleq P(V_j, U \mid A = a, Z_j = z), \qquad a \in \{0, 1\}$, and, by held-out treatment independence, define the treatment-invariant conditional distribution of the held-out

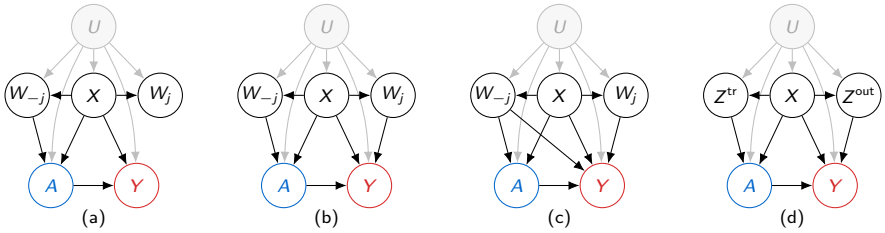



Figure 4: Causal structures compatible with proximal balancing in panels (a)–(c) and a standard two-proxy proximal structure in panel (d).

block by $K_j(\cdot \mid v,u) \triangleq P(W_j \mid V_j = v, U = u)$. For any probability distribution $Q$ on $(V_j,U)$, let $\mathcal{K}_j Q$ denote the induced distribution of $(X,W_j)$ obtained by first drawing $(V_j,U)=(X,W_{-j},U)\sim Q$, then drawing $W_j \sim K_j(\cdot \mid V_j,U)$, and finally retaining $(X,W_j)$. For each $a \in \{0,1\}$, define $\mu_{a,j,z}(v,u) \triangleq \mathbb{E}\{Y(a) \mid V_j = v, U = u, Z_j = z\}$. For $P_{Z_j}$-almost every $z$, let $\mathcal{Q}_{j,z}$ be the collection of probability distributions $Q$ on $(V_j,U)$ such that $Q \ll P(V_j,U \mid Z_j = z)$ and $\mathbb{E}_Q|\mu_{a,j,z}(V_j,U)| < \infty$ for both $a \in \{0,1\}$.

**Assumption 8** (Outcome-relevant completeness; restatement of Assumption 3)**.** For $P_{Z_j}$-almost every $z$, every $Q, Q' \in \mathcal{Q}_{j,z}$, and each $a \in \{0,1\}$, $\mathcal{K}_j Q = \mathcal{K}_j Q' \implies \mathbb{E}_Q\{\mu_{a,j,z}(V_j,U)\} = \mathbb{E}_{Q'}\{\mu_{a,j,z}(V_j,U)\}$.

Assumption 3 is stated for all pairs $Q, Q' \in \mathcal{Q}_{j,z}$ (Definition 5) within a representation stratum. The next proposition gives concrete sufficient conditions. It treats the held-out block as a measurement of $(X,U)$ that is conditionally independent of the remaining blocks, and it asks the measurement channel to be complete.

**Proposition 2** (Sufficient conditions for outcome-relevant completeness)**.** Fix a block $j$ and a representation $\phi_j$. Suppose that

(i) $W_j \perp\!\!\!\perp W_{-j} \mid (X,U)$ and, for each $a \in \{0,1\}$, $Y(a) \perp\!\!\!\perp W_{-j} \mid (X,U,Z_j)$, so that $K_j(\cdot \mid v,u)$ and $\mu_{a,j,z}(v,u)$ depend on $v=(x,w_{-j})$ only through $x$;

(ii) for every $x$, the map $\pi \mapsto \int K_j(\cdot \mid x,u)\,\pi(du)$ from probability distributions of $U$ to distributions of $W_j$ is injective.

Then Assumption 3 holds. Condition (ii) holds in the following two specializations, which are not exhaustive.

(a) *Injective decoder with additive noise.* $W_j = g(X,U) + \epsilon_j$, where $\epsilon_j$ is independent of $(X,W_{-j},U)$ and has a characteristic function without zeros, as Gaussian noise does, and $u \mapsto g(x,u)$ is injective for every $x$.

(b) *Count channel with an injective rate.* $W_j$ is a vector of counts whose coordinates are conditionally independent given $(X,U)$, with $W_{j,\ell} \mid (X=x,U=u) \sim \mathrm{Poisson}\{\lambda_\ell(x,u)\}$ for each coordinate $\ell$, and $u \mapsto \lambda(x,u) \triangleq (\lambda_1(x,u),\ldots,\lambda_d(x,u))$ is injective for every $x$.

Condition (i) lets the remaining blocks influence the outcome only through $(X,U)$ and through components retained in $Z_j$; Thm. 1, restated below, uses the more general Assumption 3. The first specialization covers latent-variable generative models in which the held-out block is a decoded image or embedding of $(X,U)$ plus independent noise, and the second covers count-valued proxies such as event, word, or sequencing-read counts. When the latent state takes finitely many values and both proxy blocks are pure measurements satisfying the rank conditions of Kuroki & Pearl (2014), their matrix restoration identifies $\tau$ directly.

### B.1.2 Exact Identification for a Fixed Block

The next theorem gives the central identification result. It states when adjustment for $Z_j$ identifies the population average treatment effect.

**Theorem 7** (Identification from a held-out audit; restatement of Thm. 1)**.** Fix a prespecified block $j \in \{1, \dots, J\}$. Suppose Assumptions 1–3 hold for this block and its representation. If $Z_j = \phi_j(V_j)$ is constructed such that $0 < P(A = 1 \mid Z_j) < 1$ almost surely, and $(X, W_j) \perp\!\!\!\perp A \mid Z_j$, then $\tau = \mathbb{E}[\mathbb{E}(Y \mid A = 1, Z_j) - \mathbb{E}(Y \mid A = 0, Z_j)]$.

The theorem is a conditional mean identification result. It establishes that the conditional mean of $Y(a)$ given $(A, Z_j)$ does not depend on $A$, and consistency converts this equality into the displayed adjustment formula. It does not assert the full conditional independence $Y(a) \perp\!\!\!\perp A \mid Z_j$.

### B.1.3 Comparison with the Canonical Two-Proxy Proximal Model

Graphically and structurally, proximal balancing differs from the standard proximal causal inference framework (Miao et al., 2018; Tchetgen Tchetgen et al., 2024). The standard framework prespecifies a treatment-inducing proxy $Z^{\mathrm{tr}}$ and an outcome-inducing proxy $Z^{\mathrm{out}}$. By contrast, proximal balancing requires only the held-out block $W_j$ to satisfy $W_j \perp\!\!\!\perp A \mid (U, V_j)$. The remaining block is not subject to this restriction and may affect $A$, $Y$, both, or neither, as illustrated by panels (a)–(c).

The underlying assumptions and identification strategies are also different. A representative canonical completeness condition is, for every $(a, x)$, $\mathbb{E}\{g(U) \mid Z^{\mathrm{tr}}, A = a, X = x\} = 0$ almost surely $\implies$ $g(U) = 0$ almost surely for functions $g$ in the specified class. Identification then proceeds by seeking an outcome bridge $h$ satisfying an observed conditional-moment equation such as $\mathbb{E}(Y \mid Z^{\mathrm{tr}}, A, X) = \mathbb{E}\{h(Z^{\mathrm{out}}, A, X) \mid Z^{\mathrm{tr}}, A, X\}$.

Proximal balancing instead combines observable balance in Thm. 1 with outcome-relevant completeness in Assumption 3. For comparison, the canonical proximal conditions refer to proxy validity, existence of an integrable outcome bridge, and treatment-proxy completeness for a designated proxy pair. The proximal-balancing conditions refer to Assumptions 1–3, representation overlap, and exact balance for a designated proxy split. A natural question is whether either framework's sufficient conditions contain those of the other.

**Proposition 3** (Non-nesting of canonical proximal and proximal-balancing sufficient conditions)**.** There exists a causal data-generating process for which the canonical proximal conditions hold for a designated proxy pair $(Z^{\mathrm{tr}}, Z^{\mathrm{out}})$ but the proximal-balancing conditions fail for the split with $W_{-j} = Z^{\mathrm{tr}}$ and $W_j = Z^{\mathrm{out}}$ under every representation $\phi_j$. Conversely, there exists a causal data-generating process for which the proximal-balancing conditions hold for a designated proxy split but the canonical proximal conditions fail for the pair $(Z^{\mathrm{tr}}, Z^{\mathrm{out}}) = (W_{-j}, W_j)$. Moreover, neither canonical treatment-proxy completeness nor Assumption 3 implies the other.

## B.2 Identification When the Valid Held-Out Block Is Unknown

Section B.1 fixes the held-out block in advance. We now allow its composition to be unknown. Consider every nonempty proper subset

$$\mathfrak{S} \triangleq \{S \subset \{1, \dots, J\} : 1 \le |S| \le J - 1\}. \tag{15}$$

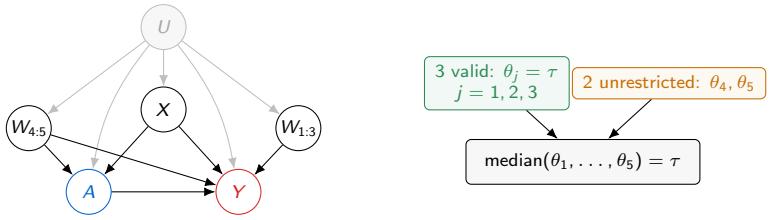
A subset $S$ and its complement $S^c$ are distinct because they exchange the held-out and remaining roles. For each $S \in \mathfrak{S}$, let $W_S = (W_j)_{j \in S}$ be the held-out vector, let $W_{-S} = (W_j)_{j \notin S}$ be the remaining vector, and define $V_S \triangleq (X, W_{-S})$, $Z_S \triangleq \phi_S(V_S)$, using the same construction as for a single held-out block.

For each $S \in \mathfrak{S}$, define, whenever it is well-defined, $\theta_S \triangleq \mathbb{E}[\mathbb{E}(Y \mid A = 1, Z_S) - \mathbb{E}(Y \mid A = 0, Z_S)]$.

Thm. 1 applies to each proxy split $S$ after replacing $(V_j, W_j, Z_j)$ by $(V_S, W_S, Z_S)$ throughout. Thus, if Assumption 1 and the grouped versions of Assumptions 2 and 3 hold, representation overlap holds, and $(X, W_S) \perp\!\!\!\perp A \mid Z_S$, then $\theta_S = \tau$.

The balance condition is observable at the population level. Define the population balance-screen set

$$\mathcal{B} \triangleq \big\{S \in \mathfrak{S} : 0 < P(A = 1 \mid Z_S) < 1 \text{ a.s.}, (X, W_S) \perp\!\!\!\perp A \mid Z_S\big\}. \tag{16}$$




(a) A majority-valid proxy structure (b) Identification by median aggregation

Figure 5: Median aggregation over five singleton proxy splits.

Each $S \in \mathcal{B}$ defines one retained proxy split and one candidate effect $\theta_S$ through Eq. (3). We assume that each such $\theta_S$ is well defined in $\mathbb{R}$. The next result considers the case in which the true effect $\tau$ occurs more often than any other candidate value.

**Theorem 8** (Identification with an unknown valid block; restatement of Thm. 2)**.** Suppose $\mathcal{B} \neq \varnothing$ and the target has a unique proximal plurality (Definition 2). Then $\{\tau\} = \arg\max_{c \in \mathbb{R}} \big|\{S \in \mathcal{B} : \theta_S = c\}\big|$.

The theorem states that $\tau$ can be identified even when we do not know the exact held-out proxy $W_j$, if it is the unique most frequent candidate value. The following corollary is the strict-majority special case of Thm. 2. It gives both the unique most frequent candidate value and the corresponding median aggregation result.

**Corollary 1** (Target-valued majority and median aggregation)**.** Suppose $\mathcal{B} \neq \varnothing$ and

$$\big|\{S \in \mathcal{B} : \theta_S = \tau\}\big| > \frac{|\mathcal{B}|}{2}. \tag{17}$$

That is, more than half of the retained proxy splits yield the true causal effect. Then $\{\tau\} = \arg\max_{c \in \mathbb{R}} \big|\{S \in \mathcal{B} : \theta_S = c\}\big|$, and $\mathrm{median}\big((\theta_S)_{S \in \mathcal{B}}\big) = \tau$.

Fig. 5 illustrates the result. The labels $W_{1:3}$ and $W_{4:5}$ denote grouped proxy blocks, not single variables. Under Assumption 1, if all five singleton proxy splits pass the balance screen and proxy splits $1, 2, 3$ satisfy the grouped versions of Assumptions 2 and 3, then three of the five singleton candidate values equal $\tau$, so their median equals $\tau$.

### B.3 Approximate Balance and Bias Bound

Exact balance is an ideal endpoint. We now consider representations that achieve only approximate, rather than exact, balance. Fix a measurable encoder $\phi$ and write $T_j \triangleq (X, W_j)$; the notation $Z_j^\phi$ was defined at the beginning of this section. Define $e_{j,\phi}(z) \triangleq P(A = 1 \mid Z_j^\phi = z), q_{j,\phi}(t, z) \triangleq P(A = 1 \mid T_j = t, Z_j^\phi = z)$. Their difference is the treatment information that remains in the held-out variables after conditioning on the representation.

**Definition 6** (Residual treatment discrepancy; restatement of Definition 3)**.** Define $D^2_{\mathrm{res},j}(\phi) \triangleq \mathbb{E}\Big[\{q_{j,\phi}(T_j, Z_j^\phi) - e_{j,\phi}(Z_j^\phi)\}^2\Big]$ and $D_{\mathrm{res},j}(\phi) \triangleq \{D^2_{\mathrm{res},j}(\phi)\}^{1/2}$.

The discrepancy measures the treatment information in $T_j$ that remains after conditioning on $Z_j^\phi$.

**Proposition 4** (Zero discrepancy; restatement of Prop. 1)**.** $D_{\mathrm{res},j}(\phi) = 0 \iff A \perp\!\!\!\perp T_j \mid Z_j^\phi$.

To relate the discrepancy with the causal bias bound, we define $P_{\phi,z} \triangleq P(T_j \mid Z_j^\phi = z)$. For a signed measure $\nu$ with a density with respect to $P_{\phi,z}$, write $\|\nu\|_{\phi,z} \triangleq \|d\nu/dP_{\phi,z}\|_{L_2(P_{\phi,z})}$. For

$P_{Z_j^\phi}$-almost every $z$ and $a \in \{0,1\}$, write $Q^\phi_{a,j,z} \triangleq P(V_j, U \mid A = a, Z_j^\phi = z)$ and $\mu^\phi_{a,j,z}(v,u) \triangleq \mathbb{E}\{Y(a) \mid V_j = v, U = u, Z_j^\phi = z\}$. Let $\mathcal{Q}^\phi_{j,z}$ be the collection of probability distributions $Q$ on $(V_j, U)$ such that $Q \ll P(V_j, U \mid Z_j^\phi = z)$ and $\mathbb{E}_Q|\mu^\phi_{a,j,z}(V_j,U)| < \infty$ for both $a \in \{0,1\}$. We assume the following:

**Assumption 9** ($L_2$ outcome-relevant stability; restatement of Assumption 4)**.** For every representation $\phi$ under consideration, there is a finite scalar $\Gamma_j(\phi) > 0$ such that, for $P_{Z_j^\phi}$-almost every $z$, every $a \in \{0,1\}$, and all $Q, Q' \in \mathcal{Q}^\phi_{j,z}$, $\left|\mathbb{E}_Q\{\mu^\phi_{a,j,z}\} - \mathbb{E}_{Q'}\{\mu^\phi_{a,j,z}\}\right| \le \Gamma_j(\phi)\, \|\mathcal{K}_j Q - \mathcal{K}_j Q'\|_{\phi,z}$.

If $\mathcal{K}_j Q = \mathcal{K}_j Q'$, the right-hand side is zero. Assumption 4 therefore implies Assumption 3 for $Z_j^\phi$.

**Theorem 9** (Approximate-balance bias; restatement of Thm. 3)**.** Suppose Assumptions 1, 2, and 4 hold for $Z_j^\phi$. Suppose there is a constant $\eta \in (0, 1/2]$ such that $\eta \le e_{j,\phi}(z) \le 1 - \eta$ for $P_{Z_j^\phi}$-almost every $z$. Then $\left|\mathbb{E}\Big[\mathbb{E}(Y \mid A = 1, Z_j^\phi) - \mathbb{E}(Y \mid A = 0, Z_j^\phi)\Big] - \tau\right| \le \frac{\Gamma_j(\phi)}{\eta(1-\eta)} D_{\mathrm{res},j}(\phi)$.

The theorem implies the following sensitivity result.

**Proposition 5** (Population sensitivity region)**.** Suppose Assumptions 1, 2, and 4 hold for $Z_j^\phi$, and suppose there is a constant $\eta \in (0, 1/2]$ such that $\eta \le e_{j,\phi}(z) \le 1 - \eta$ for $P_{Z_j^\phi}$-almost every $z$. Define $\tau_{Z_j^\phi} \triangleq \mathbb{E}\Big[\mathbb{E}(Y \mid A = 1, Z_j^\phi) - \mathbb{E}(Y \mid A = 0, Z_j^\phi)\Big]$. Then $\tau \in \left[\tau_{Z_j^\phi} - \frac{\Gamma_j(\phi)}{\eta(1-\eta)} D_{\mathrm{res},j}(\phi), \tau_{Z_j^\phi} + \frac{\Gamma_j(\phi)}{\eta(1-\eta)} D_{\mathrm{res},j}(\phi)\right]$.

This is a population sensitivity region, not a finite-sample confidence interval. When $D_{\mathrm{res},j}(\phi) = 0$, it collapses to the point-identified value in Thm. 1. The residual discrepancy is a functional of the observed distribution. The constant $\Gamma_j(\phi)$ is structural and must be justified by scientific knowledge, additional theory, or sensitivity analysis.

# C Representation Learning and AIPW Estimation

The identification results concern one fixed representation. We now select a representation from data and estimate the adjustment functional that it induces. The representation-learning and AIPW analysis below uses one prespecified held-out block $j$ throughout.

## C.1 Representation Learning

Recall that one observed record is $O_i = (X_i, W_i, A_i, Y_i)$. Partition the observed sample into three mutually disjoint subsets, $\mathcal{I}_{\mathrm{D}}$, $\mathcal{I}_{\mathrm{N}}$, $\mathcal{I}_{\mathrm{E}}$, with respective sizes $n_{\mathrm{D}}$, $n_{\mathrm{N}}$, and $n_{\mathrm{E}}$. The discrepancy-learning sample $\mathcal{I}_{\mathrm{D}}$ selects a representation. The nuisance-learning sample $\mathcal{I}_{\mathrm{N}}$ estimates the outcome regressions and propensity used by AIPW. The evaluation sample $\mathcal{I}_{\mathrm{E}}$ calculates the final estimate. Conditional on the preceding samples, the objects evaluated on the next sample are fixed.

Let $\mathcal{F}_{j,n_{\mathrm{D}}} \subseteq \Phi_j$ be a prespecified, possibly infinite class of deterministic encoders. For $\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}$, write $Z_j^\phi = \phi(V_j)$ and retain $T_j = (X, W_j)$ from Section B.3.

Assumption 5 is imposed on $\mathcal{F}_{j,n_{\mathrm{D}}}$. For any prespecified nonempty proper subset $S \subset \{1, \dots, J\}$, the results below also apply with $(V_j, W_j, Z_j^\phi)$ replaced by $(V_S, W_S, \phi(V_S))$ and with the same assumptions imposed on this grouped split.

The Brier-risk definition below estimates residual discrepancy by comparing squared-loss prediction of $A$ from $Z_j^\phi$ alone with prediction from $(T_j, Z_j^\phi)$ (Brier, 1950).

**Definition 7** (Empirical Brier discrepancy; general form of Definition 4)**.** Let $\mathcal{H}_{0,n_{\mathrm{D}}}$ and $\mathcal{H}_{1,n_{\mathrm{D}}}$ be prespecified classes of measurable $[0,1]$-valued predictors with inputs $Z_j^\phi$ and $(T_j, Z_j^\phi)$, respectively, fixed independently of $\mathcal{I}_{\mathrm{D}}$. Assume $\mathcal{H}_{1,n_{\mathrm{D}}}$ contains each member of $\mathcal{H}_{0,n_{\mathrm{D}}}$ lifted to ignore $T_j$. Define $R_{0,\phi}(h) \triangleq \mathbb{E}[\{A - h(Z_j^\phi)\}^2]$ and $\widehat{R}_{0,\phi}(h) \triangleq n_{\mathrm{D}}^{-1}\sum_{i\in\mathcal{I}_{\mathrm{D}}}\{A_i - h(Z_{j,i}^\phi)\}^2$. Likewise, define $R_{1,\phi}(h) \triangleq \mathbb{E}[\{A - h(T_j, Z_j^\phi)\}^2]$ and $\widehat{R}_{1,\phi}(h) \triangleq n_{\mathrm{D}}^{-1}\sum_{i\in\mathcal{I}_{\mathrm{D}}}\{A_i - h(T_{j,i}, Z_{j,i}^\phi)\}^2$, where $Z_{j,i}^\phi = \phi(V_{j,i})$. For each $\phi$, let $\widehat{h}_{b,\phi} \in \mathcal{H}_{b,n_{\mathrm{D}}}$ be an approximate empirical minimizer satisfying, for some $\varepsilon_{b,\phi} \geq 0$,

$$\widehat{R}_{b,\phi}(\widehat{h}_{b,\phi}) \leq \inf_{h\in\mathcal{H}_{b,n_{\mathrm{D}}}} \widehat{R}_{b,\phi}(h) + \varepsilon_{b,\phi}, \qquad b \in \{0,1\}. \tag{18}$$

Define $\varepsilon_{\mathrm{crit},j} \triangleq \sup_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}}(\varepsilon_{0,\phi} + \varepsilon_{1,\phi})$ and $\widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) \triangleq \max\Big\{(\widehat{R}_{0,\phi}(\widehat{h}_{0,\phi}) - \widehat{R}_{1,\phi}(\widehat{h}_{1,\phi})), 0\Big\}$.

Each encoder $\phi$ and critic $h$ give a squared-loss function of one observation: $\{A - h(\phi(V_j))\}^2$ or $\{A - h(T_j, \phi(V_j))\}^2$. Varying both $\phi$ and $h$ over their prespecified classes gives the loss classes $\mathcal{L}_{0,j,n_{\mathrm{D}}}$ and $\mathcal{L}_{1,j,n_{\mathrm{D}}}$, respectively. Let $\mathfrak{R}_{n_{\mathrm{D}}}(\mathcal{L})$ denote the expected Rademacher complexity of a loss class $\mathcal{L}$. For $\delta_{\mathrm{D}} \in (0,1)$, define

$$r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) \triangleq 2\{\mathfrak{R}_{n_{\mathrm{D}}}(\mathcal{L}_{0,j,n_{\mathrm{D}}}) + \mathfrak{R}_{n_{\mathrm{D}}}(\mathcal{L}_{1,j,n_{\mathrm{D}}})\} + 2\sqrt{\tfrac{\log(4/\delta_{\mathrm{D}})}{2n_{\mathrm{D}}}}.$$

This radius controls sampling fluctuation uniformly over the encoder–critic choices. The complexity terms account for the flexibility of the two loss classes, and the last term sets the failure-probability level $\delta_{\mathrm{D}}$.

Throughout, the loss classes are taken pointwise measurable and separable, so that the suprema below are measurable, and the selected encoders and critics are measurable functions of the training observations and the algorithmic seed; every statement otherwise holds with outer probability.

**Proposition 6** (Uniform residual-discrepancy deviation)**.** For each $\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}$, define

$$\begin{aligned} a_{0,n_{\mathrm{D}}}(\phi) &\triangleq \inf_{h\in\mathcal{H}_{0,n_{\mathrm{D}}}} R_{0,\phi}(h) - \mathbb{E}[\{A - e_{j,\phi}(Z_j^\phi)\}^2], \\ a_{1,n_{\mathrm{D}}}(\phi) &\triangleq \inf_{h\in\mathcal{H}_{1,n_{\mathrm{D}}}} R_{1,\phi}(h) - \mathbb{E}[\{A - q_{j,\phi}(T_j, Z_j^\phi)\}^2], \\ a_{j,n_{\mathrm{D}}} &\triangleq \sup_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} \{a_{0,n_{\mathrm{D}}}(\phi) + a_{1,n_{\mathrm{D}}}(\phi)\}. \end{aligned} \tag{19}$$

With probability at least $1 - \delta_{\mathrm{D}}$,

$$\sup_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} |\widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) - D^2_{\mathrm{res},j}(\phi)| \leq r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) + a_{j,n_{\mathrm{D}}} + \varepsilon_{\mathrm{crit},j}. \tag{20}$$

At any fixed confidence level $1-\delta_{\mathrm{D}}$, this error bound is $O(n_{\mathrm{D}}^{-1/2})$ if both Rademacher complexities, $a_{j,n_{\mathrm{D}}}$, and $\varepsilon_{\mathrm{crit},j}$ are each $O(n_{\mathrm{D}}^{-1/2})$.

**Definition 8** (Approximate empirical minimizer)**.** For a prescribed tolerance $\varepsilon_{\mathrm{enc},j} \geq 0$, an approximate empirical minimizer is an encoder $\widehat{\phi}_j \in \mathcal{F}_{j,n_{\mathrm{D}}}$ selected measurably from the training observations and algorithmic seed such that

$$\widetilde{D}^2_{j,n_{\mathrm{D}}}(\widehat{\phi}_j) \leq \inf_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} \widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) + \varepsilon_{\mathrm{enc},j}. \tag{21}$$

An exact empirical minimizer has $\varepsilon_{\mathrm{enc},j} = 0$.

**Proposition 7** (Residual-discrepancy oracle inequality)**.** Let $\delta_{\mathrm{D}} \in (0,1)$. Suppose the prespecified encoder and $[0,1]$-valued critic classes, approximate critic fits in Eq. (18) of Definition 7

hold, and $\widehat{\phi}_j$ satisfies Eq. (21). Then, with probability at least $1-\delta_{\mathrm{D}}$,

$$D^2_{\mathrm{res},j}(\widehat{\phi}_j) \le \inf_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) + 2\{r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) + a_{j,n_{\mathrm{D}}} + \varepsilon_{\mathrm{crit},j}\} + \varepsilon_{\mathrm{enc},j}. \tag{22}$$

Condition on the training observations $\{O_i : i \in \mathcal{I}_{\mathrm{D}}\}$ and any algorithmic seed used in training, which fixes $\widehat{\phi}_j$. Here and below, population expectations involving this learned encoder use a fresh observation $O = (X, W, A, Y)$ independent of the training observations and seed, with this conditioning understood. The learned representation and its population adjustment functional are

$$\begin{aligned} Z_j^{\widehat{\phi}_j} &\triangleq \widehat{\phi}_j(V_j), \qquad \widehat{Z}_j \triangleq Z_j^{\widehat{\phi}_j}, \\ \tau_{\widehat{Z}_j} &\triangleq \mathbb{E}\Big[\mathbb{E}(Y \mid A=1, \widehat{Z}_j) - \mathbb{E}(Y \mid A=0, \widehat{Z}_j)\Big]. \end{aligned} \tag{23}$$

Thus $\tau_{\widehat{Z}_j} = \tau_{Z_j^{\phi}}\big|_{\phi=\widehat{\phi}_j}$, where $\tau_{Z_j^{\phi}}$ is the population adjustment functional defined in Thm. 3. The adjustment functional is used whenever its two conditional means are absolutely integrable under the distribution of $\widehat{Z}_j$.

**Theorem 10** (Learned-representation error with an explicit radius)**.** Let $\delta_{\mathrm{D}} \in (0,1)$. Suppose the prespecified encoder and $[0,1]$-valued critic classes, approximate critic fits in Eq. (18) of Definition 7 hold, and $\widehat{\phi}_j$ satisfies Eq. (21). Suppose Assumptions 1, 2, and 4 hold for every $Z_j^{\phi}$ with $\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}$, and there is a constant $\eta \in (0, 1/2]$ such that $\eta \le e_{j,\phi}(Z_j^{\phi}) \le 1-\eta$ almost surely for every such $\phi$. Then, with probability at least $1-\delta_{\mathrm{D}}$, $|\tau_{\widehat{Z}_j} - \tau| \le \frac{\Gamma_j(\widehat{\phi}_j)}{\eta(1-\eta)}\left[\inf_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) + 2\{r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) + a_{j,n_{\mathrm{D}}} + \varepsilon_{\mathrm{crit},j}\} + \varepsilon_{\mathrm{enc},j}\right]^{1/2}$.

At fixed confidence and fixed overlap constant $\eta$, the right-hand side is $O_p(n_{\mathrm{D}}^{-1/4})$ if $\inf_{\phi\in\mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) = 0$, each of $r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})$, $a_{j,n_{\mathrm{D}}}$, $\varepsilon_{\mathrm{crit},j}$, and $\varepsilon_{\mathrm{enc},j}$ is $O(n_{\mathrm{D}}^{-1/2})$, and $\Gamma_j(\widehat{\phi}_j) = O_p(1)$. This is a conditional upper-bound rate; it is not a universal or minimax rate.

### C.2 AIPW Estimation

This section estimates the adjustment functional $\tau_{\widehat{Z}_j}$ of the learned representation with an augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994; Tsiatis, 2006) computed on the evaluation sample. Define the conditional outcome means and representation propensity

$$\begin{aligned} m_{a,j}(z) &\triangleq \mathbb{E}(Y \mid A=a, \widehat{Z}_j = z), \\ e_j(z) &\triangleq P(A=1 \mid \widehat{Z}_j = z). \end{aligned} \tag{24}$$

For the observed adjustment functional $\tau_{\widehat{Z}_j}$, the following is the AIPW score: $\mathtt{IF}_j(O) \triangleq m_{1,j}(\widehat{Z}_j) - m_{0,j}(\widehat{Z}_j) + \frac{A\{Y-m_{1,j}(\widehat{Z}_j)\}}{e_j(\widehat{Z}_j)} - \frac{(1-A)\{Y-m_{0,j}(\widehat{Z}_j)\}}{1-e_j(\widehat{Z}_j)} - \tau_{\widehat{Z}_j}$. When this score is square-integrable, it is the *efficient influence function* in the nonparametric model for $(\widehat{Z}_j, A, Y)$. Bounded outcomes together with uniform representation overlap provide one sufficient condition for square integrability. Use the nuisance-learning sample, with $\widehat{Z}_{j,i} = \widehat{\phi}_j(V_{j,i})$, to estimate the nuisance functions. Denote the estimates by $\widehat{m}_{0,j}$, $\widehat{m}_{1,j}$, and $\widehat{e}_j$.

**Definition 9** (Honest AIPW estimator)**.** For $i \in \mathcal{I}_{\mathrm{E}}$, define $\widehat{Z}_{j,i} = \widehat{\phi}_j(V_{j,i})$ and $\mathtt{UIF}_j(O_i) \triangleq \widehat{m}_{1,j}(\widehat{Z}_{j,i}) - \widehat{m}_{0,j}(\widehat{Z}_{j,i}) + \frac{A_i}{\widehat{e}_j(\widehat{Z}_{j,i})}\{Y_i - \widehat{m}_{1,j}(\widehat{Z}_{j,i})\} - \frac{1-A_i}{1-\widehat{e}_j(\widehat{Z}_{j,i})}\{Y_i - \widehat{m}_{0,j}(\widehat{Z}_{j,i})\}$. The score $\mathtt{UIF}_j$ is the uncentered plug-in version of $\mathtt{IF}_j$: it replaces the nuisance functions by their estimates and omits the centering term $\tau_{\widehat{Z}_j}$. The honest AIPW estimator is $\widehat{\tau}_j^{\mathrm{AIPW}} \triangleq \frac{1}{n_{\mathrm{E}}}\sum_{i\in\mathcal{I}_{\mathrm{E}}} \mathtt{UIF}_j(O_i)$.

The plug-in score has the doubly robust remainder structure of augmented inverse probability weighting (Chernozhukov et al., 2017; Kennedy, 2022): its conditional bias is a product of the propensity error and the outcome-regression errors, as the next proposition states exactly. For $a \in \{0,1\}$, write $r_{a,j}(z) \triangleq \frac{\widehat{m}_{a,j}(z) - m_{a,j}(z)}{a\widehat{e}_j(z) + (1-a)\{1-\widehat{e}_j(z)\}}$.

**Proposition 8** (Exact signed nuisance remainder)**.** Conditional on $\mathcal{I}_{\mathrm{D}}$ and $\mathcal{I}_{\mathrm{N}}$,

$$\mathbb{E}\{\mathtt{UIF}_j(O)\} - \tau_{\widehat{Z}_j} = \mathbb{E}\Big[\{\widehat{e}_j(\widehat{Z}_j) - e_j(\widehat{Z}_j)\}\{r_{1,j}(\widehat{Z}_j) + r_{0,j}(\widehat{Z}_j)\}\Big]. \tag{25}$$

The remainder is zero if the propensity estimator is correct. It is also zero if both conditional outcome estimators are correct.

Clip the propensity estimate so that $\eta \le \widehat{e}_j(z) \le 1-\eta$. Conditional on the two learning samples, that is, on their observations and the seeds used to fit the encoder and the nuisance functions, define for a function $g$ of $\widehat{Z}_j$ the norm $\|g\|_{2,\widehat{Z}_j} \triangleq [\mathbb{E}\{g(\widehat{Z}_j)^2 \mid \mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}\}]^{1/2}$ and the realized nuisance errors $q_{e,j} \triangleq \|\widehat{e}_j - e_j\|_{2,\widehat{Z}_j}$, $\quad q_{a,j} \triangleq \|\widehat{m}_{a,j} - m_{a,j}\|_{2,\widehat{Z}_j}$, $\quad a \in \{0,1\}$. Let $\sigma_j^2 \triangleq \mathrm{Var}\{\mathtt{UIF}_j(O) \mid \mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}\}$ denote the conditional variance of the plug-in score; it is finite whenever $\mathbb{E}(Y^2) < \infty$ and the outcome estimates are square-integrable. For $\delta_{\mathrm{E}} \in (0,1)$, define

$$\varepsilon_{\mathrm{AIPW},j} \triangleq \frac{\sigma_j}{\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E}}}} + \frac{q_{e,j}(q_{0,j} + q_{1,j})}{\eta}. \tag{26}$$

In the notation of Thm. 5, $\|\cdot\|_{2,\widehat{Z}_j} = \|\cdot\|_{2,j}$ and $q_{e,j}(q_{0,j} + q_{1,j})/\eta = \mathtt{err}_j$, so $\varepsilon_{\mathrm{AIPW},j} = \sigma_j/\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E}}} + \mathtt{err}_j$. The last term of the second bound of Thm. 11 below is $\mathtt{bias}_j$. The next theorem is the final guarantee of this section: the causal error of the honest AIPW estimator is at most the estimation radius $\varepsilon_{\mathrm{AIPW},j}$ plus the stability-weighted residual discrepancy of the learned representation.

**Theorem 11** (Honest AIPW error; general form of Thm. 5)**.** Suppose $\sigma_j < \infty$. Conditional on the two learning samples, with probability at least $1 - \delta_{\mathrm{E}}$ over $\mathcal{I}_{\mathrm{E}}$,

$$|\widehat{\tau}_j^{\mathrm{AIPW}} - \tau_{\widehat{Z}_j}| \le \varepsilon_{\mathrm{AIPW},j}. \tag{27}$$

If, in addition, Assumptions 1, 2, and 4 hold for $Z_j^{\widehat{\phi}_j}$ and $\eta \le e_j(\widehat{Z}_j) \le 1-\eta$ almost surely, then on the same event $|\widehat{\tau}_j^{\mathrm{AIPW}} - \tau| \le \varepsilon_{\mathrm{AIPW},j} + \frac{\Gamma_j(\widehat{\phi}_j)}{\eta(1-\eta)} D_{\mathrm{res},j}(\widehat{\phi}_j)$.

**Corollary 2** (Consistency)**.** Consider a sequence of encoder and critic classes, sample splits, nuisance estimators, optimization tolerances, and selected encoders $\widehat{\phi}_{j,n}$ indexed by the total sample size $n$, with a common overlap constant $\eta \in (0, 1/2]$, such that Thm. 5 applies at every $n$. Suppose the following three conditions hold.

(C1) *Vanishing evaluation error.* $\delta_{\mathrm{E},n} \to 0$ and $\sigma_{j,n}^2/(n_{\mathrm{E}}\delta_{\mathrm{E},n}) \to 0$. The evaluation sample grows faster than the confidence level shrinks.

(C2) *Vanishing nuisance product.* $q_{e,j,n}(q_{0,j,n} + q_{1,j,n}) \xrightarrow{p} 0$. The propensity and outcome-regression errors need not vanish individually; their product vanishes, for example, when each is $o_p(n^{-1/4})$.

(C3) *Vanishing learned imbalance.* $\Gamma_j(\widehat{\phi}_{j,n}) D_{\mathrm{res},j}(\widehat{\phi}_{j,n}) \xrightarrow{p} 0$. By Prop. 7, this holds when $\inf_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} D_{\mathrm{res},j}^2(\phi)$, the two Rademacher complexities, $a_{j,n_{\mathrm{D}}}$, $\varepsilon_{\mathrm{crit},j,n}$, $\varepsilon_{\mathrm{enc},j,n}$, $\delta_{\mathrm{D},n}$, and $\log(1/\delta_{\mathrm{D},n})/n_{\mathrm{D}}$ all converge to zero and $\Gamma_j(\widehat{\phi}_{j,n}) = O_p(1)$.

Then $\widehat{\tau}_{j,n}^{\mathrm{AIPW}} \xrightarrow{p} \tau$.

Eq. (13) separates the final error into an AIPW estimation term and a hidden-confounding sensitivity term. The selected stability constant $\Gamma_j(\widehat{\phi}_j)$, critic approximation error $a_{j,n_{\mathrm{D}}}$, optimization tolerances, and nuisance errors are not generally observed. The result becomes an operational numerical certificate after these quantities are bounded by theory, computation, external information, or a declared sensitivity analysis.

**Algorithm 2:** PROBE split search, detailed form

**Input:** a block partition $W = (W_1, \dots, W_J)$; observations split into $\mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}, \mathcal{I}_{\mathrm{E}}$; a budget $m$; a threshold $t \geq 0$; an overlap constant $\eta$; a linking radius $\rho > 0$.

**Output:** an estimate $\widehat{\tau}$ of $\tau$, or failure.

1 Draw $m$ proxy splits uniformly without replacement from $\mathfrak{S}$;

2 **foreach** *sampled split* $S$ **do**

3 learn $\widehat{\phi}_S$ on $\mathcal{I}_{\mathrm{D}}$ by minimizing $\widetilde{D}^2_{S,n_{\mathrm{D}}}$;

4 fit an unclipped propensity $\widehat{e}_S^{\mathrm{raw}}$ on $\mathcal{I}_{\mathrm{N}}$; retain $S$ if $\widetilde{D}^2_{S,n_{\mathrm{D}}}(\widehat{\phi}_S) \leq t$ and $\widehat{e}_S^{\mathrm{raw}}$ passes a prespecified empirical overlap check at level $\eta$.;

5 Let $\mathcal{R}_m$ be the set of retained splits;

6 **foreach** $S \in \mathcal{R}_m$ **do**

7 fit outcome regressions on $\mathcal{I}_{\mathrm{N}}$, set $\widehat{e}_S = \mathrm{clip}(\widehat{e}_S^{\mathrm{raw}}, \eta, 1-\eta)$, and compute the honest AIPW estimate $\widehat{\theta}_S$ from Eq. (10) on $\mathcal{I}_{\mathrm{E}}$.;

8 Link $S$ and $S'$ whenever $|\widehat{\theta}_S - \widehat{\theta}_{S'}| \leq 2\rho$, and form the connected components;

9 **return** the median of $\{\widehat{\theta}_S\}$ over the unique largest component; declare failure if several components share the largest size;

### C.3 Searching Across Proxy Splits

Exhaustive evaluation of all proxy splits in $\mathfrak{S}$, defined in Eq. (15), may be computationally expensive. The number of splits, $|\mathfrak{S}| = 2^J - 2$, grows exponentially in the number of blocks $J$, and each split requires its own representation and nuisance fits.

Algorithm 2 therefore samples $m$ proxy splits instead of evaluating all of them, and then works in four stages. The sampling stage draws the splits $S$ (line 1). The screening stage keeps the splits that pass the empirical balance screen (lines 2–5). The estimation stage computes one honest AIPW estimate per retained split (lines 6–7). The aggregation stage links estimates within distance $2\rho$ and returns the median of the largest component (lines 8–9). The theorem below gives conditions under which those empirical components recover a fixed oracle partition of the population candidates.

Three questions decide whether the output is accurate: whether enough screened splits are sampled, whether the designated target cell remains the unique plurality, and whether the estimated graph recovers the fixed oracle cells.

For the analysis only, preassign a learning seed $\omega_S$ to every $S \in \mathfrak{S}$ before the split draw, let $\mathcal{G} \triangleq \sigma(\mathcal{I}_{\mathrm{D}}, \mathcal{I}_{\mathrm{N}}, \{\omega_S : S \in \mathfrak{S}\})$, and condition on $\mathcal{G}$ throughout this subsection. Counterfactually applying the fixed learning-and-screening rule to every $S \in \mathfrak{S}$ defines $\widehat{\mathcal{B}}$, although Algorithm 1 evaluates only sampled splits. Thus, for every $S \in \mathfrak{S}$, the learned encoder $\widehat{\phi}_S$, the nuisance estimates, the screen outcome, and the set $\widehat{\mathcal{B}}$ are fixed. For $S \in \widehat{\mathcal{B}}$, let $\widehat{Z}_S \triangleq \widehat{\phi}_S(V_S)$ and let $\theta_S \triangleq \tau_{\widehat{Z}_S}$ be its population adjustment value.

Let $\Pi = \{C_0, \dots, C_L\}$ be a partition of $\widehat{\mathcal{B}}$ that is fixed given $\mathcal{G}$ and hence chosen before the random split draw and $\mathcal{I}_{\mathrm{E}}$. Because the population values $\theta_S$ are generally unknown, $\Pi$ is an oracle device for analysis rather than an object computed by Algorithm 1.

Let $T \triangleq |\mathfrak{S}| = 2^J - 2$ and $N \triangleq |\widehat{\mathcal{B}}| \geq 1$. Fix $m \in \{1, \dots, T\}$. Line 1 of Algorithm 2 draws $m$ proxy splits uniformly without replacement from $\mathfrak{S}$, independently of $\mathcal{G}$. Let $\mathcal{R}_m$ be the sampled splits in $\widehat{\mathcal{B}}$, let $R \triangleq |\mathcal{R}_m|$, and let $R_\ell \triangleq |\mathcal{R}_m \cap C_\ell|$. Then

$$R \sim \mathrm{Hypergeometric}(T, N, m), \tag{28}$$

since exactly $N$ of the $T$ candidates belong to $\widehat{\mathcal{B}}$ and the $m$ draws are made without replacement.

For the oracle cells, define

$$n_\ell \triangleq |C_\ell|, \qquad \ell = 0, \dots, L. \tag{29}$$

Adopt the convention that a maximum over an empty collection equals zero and define

$$\Delta_\Pi \triangleq \frac{n_0}{N} - \max_{1\le \ell \le L} \frac{n_\ell}{N}. \tag{30}$$

Thus $\Delta_\Pi > 0$ means that the target cell $C_0$ is larger than every competing oracle cell. If $L = 0$, the maximum in Eq. (30) and the retained-cell maximum in Lemma 1 are both defined as zero.

Two quantities of the distribution in Eq. (28) enter the guarantee:

$$\begin{aligned} P(R=0) &= \binom{T-N}{m} \Big/ \binom{T}{m}, \\ \mathbb{E}\{e^{-R\Delta_\Pi^2/2}\} &= \sum_{r=0}^{\min(m,N)} \binom{N}{r}\binom{T-N}{m-r} e^{-r\Delta_\Pi^2/2} \Big/ \binom{T}{m}. \end{aligned} \tag{31}$$

Eq. (31) gives the empty-draw probability and the plurality-reversal term; both decrease as the budget $m$ grows.

**Lemma 1** (Cluster plurality under random subsampling)**.** Conditional on $\mathcal{G}$, $R \sim \mathrm{Hypergeometric}(T, N, m)$. If $\Delta_\Pi > 0$, then with probability at least $1 - P(R = 0) - L\,\mathbb{E}\{e^{-R\Delta_\Pi^2/2}\}$, $R \ge 1$ and $R_0 > \max_{1\le \ell \le L} R_\ell$, with an empty maximum equal to zero.

Fix $\delta \in (0, 1)$. We call $\Pi$ $(b, \varepsilon, \rho, \delta)$-recoverable if $b, \varepsilon \ge 0$, $b + \varepsilon \le \rho$, $|\theta_S - \tau| \le b$ for every $S \in C_0$, within-cell distances are at most $2(\rho - \varepsilon)$, between-cell distances exceed $2(\rho + \varepsilon)$, and, conditional on $\mathcal{G}$, $|\widehat{\theta}_S - \theta_S| \le \varepsilon$ for every $S \in \widehat{\mathcal{B}}$ with probability at least $1 - \delta$ over $\mathcal{I}_\mathrm{E}$.

The following theorem bounds the error of the output of Algorithm 1. Its probability is over the evaluation sample and the random draw of splits, conditional on the learning samples.

**Theorem 12** (Finite-sample PROBE cluster recovery; general form of Thm. 6)**.** Conditional on $\mathcal{G}$, suppose $\Delta_\Pi > 0$ and $\Pi$ is $(b, \varepsilon, \rho, \delta)$-recoverable. Then, over the independent split draw and $\mathcal{I}_\mathrm{E}$, the output $\widehat{\tau}$ of Algorithm 1 run with linking radius $\rho$ satisfies $P\big(|\widehat{\tau} - \tau| \le b + \varepsilon\big) \ge 1 - \delta - P(R = 0) - L\,\mathbb{E}\big\{e^{-R\Delta_\Pi^2/2}\big\}$, where $R \sim \mathrm{Hypergeometric}(T, N, m)$ as in Eq. (28), the two sampling terms are given in Eq. (31), and a declared failure counts as $|\widehat{\tau} - \tau| > b + \varepsilon$.

The subsampling lemma controls whether $C_0$ remains the unique largest retained cell. The within-cell and between-cell conditions then make the graph components in line 8 of Algorithm 2 equal the retained restrictions of the oracle cells. The target-band and evaluation-error conditions place every estimate in the selected target component within $b + \varepsilon$ of $\tau$. Eq. (27) at $\delta_\mathrm{E} = \delta/N$ and a union bound over $\widehat{\mathcal{B}}$ supply $\varepsilon = \max_{S\in\widehat{\mathcal{B}}} \varepsilon_{\mathrm{AIPW},S}$, and Eq. (6) supplies $b = \max_{S\in C_0} \Gamma_S(\widehat{\phi}_S) D_{\mathrm{res},S}(\widehat{\phi}_S)/\{\eta(1-\eta)\}$ when every $S \in C_0$ satisfies Assumptions 1, 2, and 4; this gives Thm. 6 of the main text, with $\Pi = \{C_0, \ldots, C_L\}$ and $\Delta = \Delta_\Pi$. The partition and its gap remain oracle quantities and are not certified by the empirical graph. When $b = 0$ and $C_0 = \{S : \theta_S = \tau\}$, the target-band premise uses the exact population equality notion of Section 3.2; the finite-sample theorem additionally involves empirical screening, random subsampling, and estimation error.

The evaluation failure $\delta$ is paid with the uniform radius $\varepsilon$, while the two sampling terms are paid with the budget $m$. The fraction $p \triangleq N/T$, the oracle margin $\Delta_\Pi$, and the number $L$ of competing cells are fixed after conditioning. The next corollary converts a sampling-failure tolerance $\alpha$ into a budget.

**Corollary 3** (Budget for a target failure level)**.** Condition on $\mathcal{I}_\mathrm{D}$, $\mathcal{I}_\mathrm{N}$, and the learning seeds. Let $T \triangleq |\mathfrak{S}|$, let $\widehat{\mathcal{B}}$ be a fixed screened set of size $N \ge 1$, and define $\theta_S \triangleq \tau_{\widehat{\phi}_S(V_S)}$. Before the split draw and $\mathcal{I}_\mathrm{E}$, fix a partition $\Pi = \{C_0, \ldots, C_L\}$ of $\widehat{\mathcal{B}}$ that is fixed given $\mathcal{G}$. Set $n_\ell \triangleq |C_\ell|$, $\Delta_\Pi \triangleq n_0/N - \max_{1\le \ell \le L} n_\ell/N > 0$, and $p \triangleq N/T$, with an empty maximum equal to zero. Draw $m \in \{1, \ldots, T\}$ splits uniformly without replacement, let $R$ be the number drawn from

$\widehat{\mathcal{B}}$, and fix $\alpha \in (0,1)$. Then

$$P(R=0) + L\,\mathbb{E}\{e^{-R\Delta_\Pi^2/2}\} \;\le\; (L+1)\exp\{-m\,p\,(1-e^{-\Delta_\Pi^2/2})\}. \tag{32}$$

Suppose additionally that $b, \varepsilon \ge 0$, $b+\varepsilon \le \rho$, $|\theta_S - \tau| \le b$ on $C_0$, within-cell distances are at most $2(\rho-\varepsilon)$, between-cell distances exceed $2(\rho+\varepsilon)$, and $|\widehat{\theta}_S - \theta_S| \le \varepsilon$ simultaneously with probability at least $1-\delta$ over $\mathcal{I}_{\mathrm{E}}$. With every largest-component tie counted as failure, Algorithm 1 then satisfies $P(|\widehat{\tau}-\tau| \le b+\varepsilon) \ge 1-\delta-\alpha$ whenever $m \ge \frac{\log\{(L+1)/\alpha\}}{p\,(1-e^{-\Delta_\Pi^2/2})}$, and the simpler budget $m \ge 3\log\{(L+1)/\alpha\}/(p\Delta_\Pi^2)$ is sufficient. If $L=0$, the budget $m \ge \log(1/\alpha)/p$ suffices.

The budget scales as $\log\{(L+1)/\alpha\}/(p\Delta_\Pi^2)$. The ratio $R/m$ is unbiased for $p$, but $\Delta_\Pi$ is defined by the oracle partition and is not identified by the realized graph without additional assumptions.

**Corollary 4** (Median under majority)**.** Condition on the learning samples and seeds. Let $\widehat{\mathcal{B}}$ be fixed, define $\theta_S \triangleq \tau_{\widehat{\phi}_S(V_S)}$, and before the split draw and $\mathcal{I}_{\mathrm{E}}$ fix a partition of $\widehat{\mathcal{B}}$ that is fixed given $\mathcal{G}$ and contains a target cell $C_0$. Let $\mathcal{R}_m \subseteq \widehat{\mathcal{B}}$ be the retained splits with $R \triangleq |\mathcal{R}_m|$. Fix $b, \varepsilon \ge 0$. Suppose $|\theta_S - \tau| \le b$ for every $S \in C_0$, and let $E$ be an event on which $|\widehat{\theta}_S - \theta_S| \le \varepsilon$ for every $S \in \mathcal{R}_m$. If $|\mathcal{R}_m \cap C_0| > R/2$, then on $E$ the median of $\{\widehat{\theta}_S : S \in \mathcal{R}_m\}$ lies within $b+\varepsilon$ of $\tau$. The linking step and cell-separation conditions are not needed.

When the target cell forms a strict retained majority, Cor. 4 dispenses with the linking step. With $b=0$ and $C_0 = \{S : \theta_S = \tau\}$, its target-band premise reduces to the exact population target class used in Cor. 1.

# D Experiment Details and Evidence Boundaries

This appendix records the technical details behind Section 5. The tables in this appendix report the original runs of each experiment. Fig. 3 and its detailed versions, Figs. 8 and 9, also include data sets added later with seeds fixed in advance, for 100 data sets per synthetic setting, 100 Twins replicates, and 88 IHDP replicates in total, so their MAE values differ from the tables.

## D.1 Shared structural family

All four confirmations share one structure. Let $U$ be the hidden confounder, $I$ a treatment-only cause, $C$ an observed confounder, and let $\eta$, $\epsilon_0, \ldots, \epsilon_4$ and $\varepsilon_Y$ be independent standard normal noises. The informative coordinates, which no learner sees by name, are $X_1 = U + 2\eta$, $M_j = U + 0.85\,\epsilon_j$ $(j = 0, \ldots, 3)$, $M_4 = U - 3.0\,I + 0.85\,\epsilon_4$. Treatment and outcome follow $P(A = 1 \mid U, I, C, X_1) = 0.1 + 0.8\,\Phi(1.5U + 2.0I + 0.2X_1 + 0.3C)$, $Y(a) = a - 2.5U + 0.2X_1 - 0.5C + 0.3\varepsilon_Y$, so the average treatment effect is $\tau = 1$. The observed covariate $X$ carries $X_1$ and $C$. Blocks $W_1, W_2, W_3$ carry $M_1, M_2, M_3$, block $W_4$ carries the contaminated $M_4$, and block $W_5$ carries $M_0$ together with $I$. Each observed vector is a fixed random rotation of its informative coordinates stacked on pure noise coordinates with standard deviation 0.5, so no observed coordinate is a named quantity. Placing $C$ inside $X$, which always enters the representation, prevents a representation from dropping an outcome-relevant coordinate in order to balance.

The three numeric settings differ only in how many noise coordinates hide the informative ones. SCM-1 gives $X$ two coordinates and the five blocks $(1, 1, 1, 1, 2)$ coordinates, for 8 in total. SCM-2 gives $X$ eight and each block sixteen, for 88. SCM-3 gives $X$ thirty-two and each block 256, for 1,312. Because the coefficients are shared, the population values and the census of balanceable splits are the same objects across the three.

In the population of this family, seven of the thirty candidate splits admit exact balance and target $\tau$, namely those built from the clean blocks $W_1, W_2, W_3$; the split that holds out $W_4$ also admits balance and targets a different value; the seven splits that mix $W_4$ with a clean block admit no balance; every split containing $W_5$ admits no balance at all. This is the census the screen is measured against when we report true and false positive rates.

### D.2 IMAGE RENDERING

SCM-4 keeps the structure above and changes only what the learner receives in place of the six observed proxy coordinates. Each coordinate is recorded to 15 equally spaced levels on $[-3\sigma, 3\sigma]$, where $\sigma$ is its exact population standard deviation, and is then shown as one original $64 \times 64$ RGB Shapes3D image whose orientation factor equals that level. Every other generative factor of the image is drawn uniformly and independently, so the image carries the recorded level and nothing else about the unit. Blocks $W_1$ to $W_4$ contribute one image each and $W_5$ contributes two, giving six images per unit.

A frozen reader maps one image back to one number. It is a convolutional network trained with orientation labels on an image bank disjoint from every image used in the causal data, and it is never refitted during the experiment. Across the $14{,}400{,}000$ reads of all 100 data sets it produced no level error, so SCM-4 measures the pipeline under an image proxy and not the discovery of a representation from raw pixels. The pixel CNN comparator in Section 5 is a convolutional network fitted on all six images without holding out a block and without the balance screen.

Two of the twenty SCM-4 data sets reached the wall-clock cap for that comparator while three tasks shared one GPU, and the sealed summary therefore recorded FAIL on the completeness gate and on the comparison that depends on it. The two comparator runs were repeated alone under the sealed configuration and the sealed seeds, and nothing else was rerun or changed. The sealed summary and the amended summary are both retained; the amended one passes all 13 gates.

### D.3 SEARCH FAMILY, SCREEN, AND AGGREGATION

The search ranges over the thirty nonempty proper subsets of the five blocks. For a split $S$ the representation is $Z_S = (X, Bv_{-S})$, where $v_{-S}$ stacks the preprocessed kept blocks and $B$ has rank $k = 2$. Preprocessing standardises $X$ and reduces any block with more than two coordinates to its principal directions above the Marchenko–Pastur edge, at most four, fitted on the representation fold alone. The family is therefore linear in the observed coordinates and role-blind: it receives no label identifying a block, but it is not an unrestricted neural encoder.

Each data set is split into four folds of nearly equal size, $6{,}000$ rows each here, rotated so that every fold serves once as representation, screen, nuisance, and evaluation sample. The critics are bounded probit models with link $0.1 + 0.8\Phi$ and ridge $10^{-4}$; the augmented critic nests the base one, and the lifted base is a candidate whenever the augmented fit cannot improve on it. The screen statistic is the held-out Brier risk difference on the screen fold, refitted there rather than reused from the search and cross-fitted by row parity. A split is retained when $\max(\mathrm{gap}, 0) + z\,\mathrm{SE}$ falls below the threshold and every base-critic prediction on that fold lies in $[0.05, 0.95]$, with $z = \Phi^{-1}(1 - 0.05/30) \approx 2.935$ for the thirty simultaneous decisions. The thresholds are 1.5, 1.8, 1.6 and 1.7 times $10^{-3}$ for SCM-1 to SCM-4, each fixed on development data before the confirmation seeds were generated. Retained splits are grouped by single linkage at twice a common radius, and the reported estimate is the median of the largest agreeing group, with no estimate returned when two groups tie for largest. Relative to Algorithm 1, this implementation searches all 30 splits ($m = T$), adds a separate screen fold to the three samples, retains a split only if it passes the screen in all four rotations, averages its estimate over the rotations, and sets the linking radius $\rho$ from the data by a Gaussian multiplier bootstrap. Appendix D.5 states every one of these choices in full.

### D.4 HONEST LEARNING AND EVALUATION

The four folds of a data set carry four separate jobs, and every rotation assigns each fold to each job once. The representation and its search critics use the representation fold. The screen refits fresh critics on the screen fold, so no statistic that decides retention was fitted on the rows that produced the representation. Propensity and outcome nuisance functions use the nuisance fold, and the AIPW score uses the evaluation fold. Preprocessing parameters are estimated only from the fold that is training at that step, and the four rotations are averaged.

For a fixed representation, the base critic predicts $A$ from $Z$, and the augmented critic predicts $A$ from $(X, W_S, Z)$. The empirical screen uses the clipped difference between their paired Brier risks. The runs additionally record critic reliability diagnostics, while the overlap check is part of the

retention rule. These diagnostics are empirical guards. They are not upper confidence bounds on the population discrepancy.

The reported paired comparison uses one value per independently generated data set. For method $M$ and baseline $B$, define $d_k = |\widehat{\tau}_{M,k} - 1| - |\widehat{\tau}_{B,k} - 1|$. The prespecified one-sided upper bound is $\overline{d} + t_{0.95,K-1} \frac{\mathrm{sd}(d_1,\ldots,d_K)}{\sqrt{K}}$. The data set, not a proxy split, row, or training restart, is the replication unit. A no-return data set is included in the return-rate denominator and is not filled with zero error.

### D.5 Modeling choices and implementation

**Why this representation family.** Two requirements shaped the choice. The screen must certify balance from a finite sample, which asks for an encoder class whose uniform deviation term in Eq. (22) is small, and the search must stay role-blind, so the class may not use any label that identifies a block. A rank-two linear map on preprocessed kept blocks meets both, and every fit it produces is deterministic given its seed, which the sealing protocol requires. The price is that this family cannot represent a nonlinear measurement channel. SCM-4 pays that price explicitly: a frozen network reads each image into a number before PROBE sees it, so the linear family acts on recovered numbers rather than on pixels.

**Preprocessing.** All preprocessing is fitted on the representation fold alone. The covariates $X$ are standardised. A block with at most two coordinates passes through unchanged. A block with more coordinates is reduced by a singular value decomposition: with eigenvalues $s^2/n$ of the centred block, the noise level is the median eigenvalue $\sigma^2$, the Marchenko–Pastur edge is $\sigma^2(1 + \sqrt{d/n})^2$ for $d$ coordinates, and the retained directions are those whose eigenvalue exceeds $1.1$ times that edge, at least one and at most four. The retained scores are standardised again.

**Representation.** For a split $S$, let $v_{-S}$ stack the preprocessed kept blocks, of width $q$, and let $B \in \mathbb{R}^{k \times q}$ with $k = \min(2, q)$, so that $Z_S = (X, v_{-S}B^\top)$. The rows of $B$ are normalised inside the objective, which makes the parameterisation scale invariant rather than constrained. The objective is the ridge-penalised Brier gap $G(B) = \widehat{R}_0 - \widehat{R}_1$ of the two nested critics on the representation fold, and its gradient follows from the envelope theorem, projected onto the unit-row tangent space. The optimiser is L-BFGS-B with at most $60$ outer iterations, tolerances $10^{-14}$ and $10^{-12}$, and two starts: the top-$k$ right singular vectors of the centred kept features, and one Gaussian random matrix. The start with the lower $G$ is kept, and the critic fits are warm-started across outer iterations. No penalty is placed on $B$.

**Critics.** Both critics are bounded probit models, $q = 0.1 + 0.8\,\Phi(\eta)$, fitted by L-BFGS-B with at most $500$ iterations on the ridge-penalised Brier risk with penalty $10^{-4}$ on every coefficient except the intercept. The bounded link is what makes Assumption 5 hold by construction for the fitted propensity and keeps the critic class $[0,1]$-valued, as Definition 4 requires. The base critic reads $(1, X, Z_S)$ and the augmented critic reads those inputs together with $T_S$. The base critic lifted into the augmented space, by padding its coefficients with zeros, is always a candidate for the augmented fit, so the empirical gap is nonnegative by construction.

**Screen.** On the screen fold the critics are refitted from scratch and cross-fitted by row parity: even rows fit and odd rows score, then the roles are reversed. The statistic is the pooled out-of-fold paired Brier difference, with $\mathrm{gap}$ its mean and $\mathrm{SE}$ its standard deviation divided by $\sqrt{n_{\mathrm{S}}}$. A split is retained when $\max(\mathrm{gap}, 0) + z\,\mathrm{SE}$ falls below the threshold and every base-critic prediction on that fold lies in $[0.05, 0.95]$, and it is retained overall only if it passes in all four rotations. The overlap condition is part of the rule, not a diagnostic.

**Aggregation.** The common radius $\rho$ comes from a Gaussian multiplier bootstrap rather than from resampling. The evaluation-fold scores of the thirty splits are stacked into one matrix, $\Sigma$ is their centred cross-product divided by $n^2$, and $\rho$ is the 95th percentile of $\max_j |\xi_j|$ over $5{,}000$ draws $\xi \sim N(0, \Sigma)$. Two splits are linked when their estimates differ by at most $2\rho$, groups are the connected components under single linkage, and the reported estimate is the median of the largest group. When two groups tie for largest, the run returns no estimate, and that data set enters the return-rate denominator.

**Nuisance functions and the AIPW score.** The propensity is the same bounded probit at ridge $10^{-4}$, fitted on the nuisance fold and clipped to $[0.05, 0.95]$. Each arm's outcome regression is a ridge least squares fit on the nuisance fold with an unpenalised intercept and penalty $10^{-4}n$. The score is evaluated on the evaluation fold, and a split's estimate is the plain mean of its four rotations.

**Folds by setting.** Every setting splits its rows into four folds of nearly equal size and rotates them, so the 6,000-row folds are specific to the synthetic settings.

| Setting | $n$ | Data sets | Fold sizes | Screen rule |
|---|---|---|---|---|
| SCM-1 to SCM-4 | 24,000 | 20 sealed, 100 total | 6,000 each | fixed threshold |
| Twins | 11,984 | 10 sealed, 100 total | 2,996 each | noise floor |
| IHDP | 747 | 10 sealed, 88 total | 187, 187, 187, 186 | noise floor |
| ACIC | 4,802 | 9, exploratory | 1,201 or 1,200 | noise floor |
| RHC | 5,735 | 1, effect unknown | 1,434 or 1,433 | noise floor |
| Kernel benchmark | 3,000 | 20 sealed, 100 total | 750 each | pooled gap |

The synthetic settings require a pass in all four rotations at the fixed threshold. The kernel benchmark disables the fixed threshold and retains a split when its pooled gap is at most $z\,\mathrm{SE}$ with $z = \Phi^{-1}(1 - 0.05/2) = 1.96$, since only two splits are available there. The noise floor pools the four rotations, taking $\mathrm{gap} = \max(\mathrm{mean}, 0)$ and $\mathrm{SE} = \sqrt{\sum \mathrm{SE}^2}/4$, and retains a split when $\mathrm{gap} + z\,\mathrm{SE}$ is at most the median of $z\,\mathrm{SE}$ over the thirty splits of the same data set. The realised floors are $0.002661$ for Twins, $0.021346$ for IHDP, $0.004694$ for ACIC, and $0.006335$ for RHC. Overlap is not part of this rule.

**The image reader and the pixel comparator.** The image bank is split once, under a fixed seed, into 200,000 reader-training images, 40,000 selection images, and 240,000 images reserved for the causal data, and the three parts are disjoint. The reader is four convolutions, $3 \to 32 \to 64 \to 128 \to 128$ channels, each $3 \times 3$ with stride 2 and padding 1 and a ReLU, followed by a flattening and two linear layers of widths 256 and 1. It is trained as a regression on the scaled level $(\ell - 7)/7$ under squared loss for six epochs with batch size 256 and Adam at learning rate $10^{-3}$, the epoch with the lowest validation error is kept, and the weights are then frozen, hashed, and reloaded in evaluation mode for every run. The pixel comparator starts from the reader: it loads the reader's convolutional tower and its 256-dimensional feature layer, adds a linear map to sixteen outputs whose first row is copied from the reader's output layer, and shares that tower across the six images. The sixteen outputs per image and the standardised covariates enter two hidden layers of widths 128 and 64 with ReLU activations and three heads for the propensity and the two outcome regressions. Training minimises the Brier loss for treatment plus the squared loss for the standardised outcome, with Adam at $10^{-4}$ for the tower and $10^{-3}$ elsewhere, five epochs, batch size 128, a validation share of 0.2, and patience two. The comparator uses PROBE's folds and rotations and the same clipping, and it carries a wall-clock cap of 2,400 seconds for its four rotations.

**Which comparators share the folds.** Adjustment for $X$, adjustment for $(X, W)$, the unchecked summary, the oracle, and the pixel comparator all use PROBE's folds and rotations, with the bounded probit propensity and arm-wise ridge outcome models whose ridge levels are chosen by three-fold cross-validation on the nuisance fold. The oracle adjusts for the latent coordinates rather than for the rotated observed covariates, so it is stronger than an estimator that sees $X$ and $U$ in observed form. CEVAE, proximal 2SLS, single-proxy control, and KSPC are fitted on the whole data set without PROBE's folds and without cross-fitting, as in their source implementations.

**Seeds and sealing.** A sealed run writes a protocol file once, recording the sample size, the full design, and a hash of every source file it will execute; a later run aborts if any hash or any design field differs. Shards are written atomically and never over an existing file, and summaries are also write-once. Fold seeds, representation start seeds, bootstrap seeds, and comparator seeds are fixed offsets of the task seed, so a rerun of a task reproduces its fits exactly on the same hardware.

### D.6 Additional results for the synthetic analysis

This subsection details the original sealed confirmation runs, summarized in Table 2. Fig. 6 shows every method on its own row, and Fig. 7 shows the kernel single-proxy benchmark.

Methods that need proxy roles receive the roles of the data-generating process in their correct-role runs. Proximal 2SLS uses $W_4$, which the treatment-only cause $I$ affects, as the treatment-side proxy, a role $W_4$ can fill although it is not a valid held-out block, and $W_2$ as the outcome-side proxy, and single-proxy control and KSPC use the clean block $W_1$. The wrong-role runs swap the two 2SLS proxies and give single-proxy control and KSPC the contaminated block $W_4$. KSPC is our implementation of the estimator of Xu & Gretton (2025) with observed covariates, using its SKPV estimator only, on a random subsample of 3,000 rows when a data set is larger.

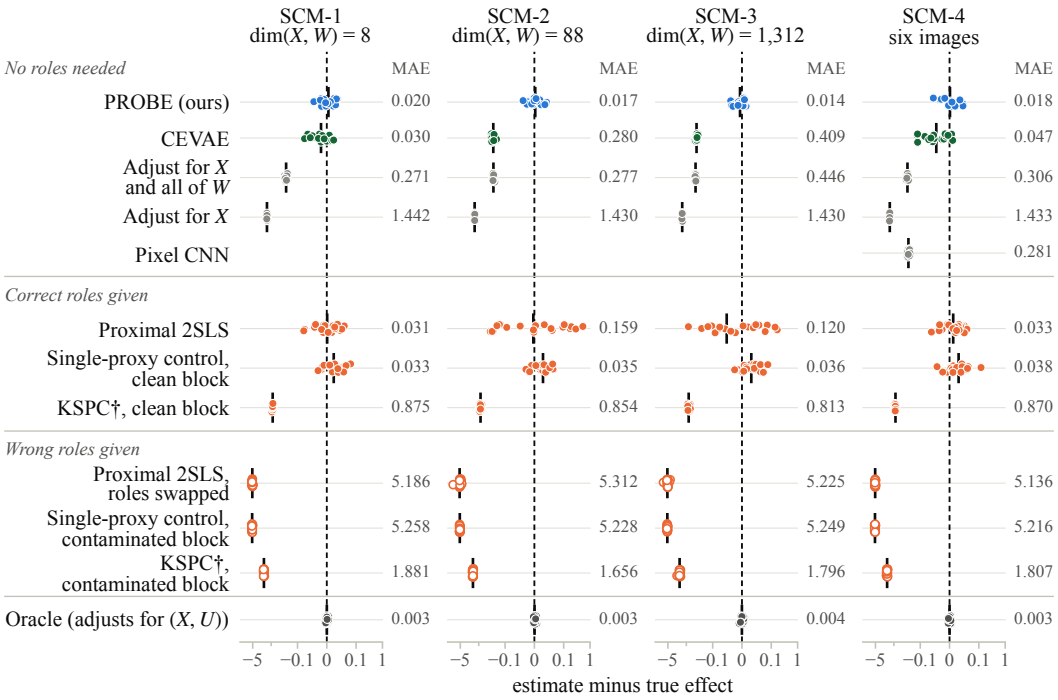
**Comparators.** Adjustment for $X$, adjustment for $(X, W)$, and the oracle, which also adjusts for $U$, are AIPW estimators on the same folds and rotations as PROBE. Proximal 2SLS is linear. It regresses the outcome-side proxy on $A$, the treatment-side proxy, and $X$, and then regresses $Y$ on $A$, the fitted proxy, and $X$; a block with more than four coordinates enters through its first four principal components. Single-proxy control fits the linear outcome bridge of Park et al. (2024) in $X$ and the first principal component of its block on the untreated units, which gives the effect on the treated. This effect equals $\tau$ in SCM-1 to SCM-4 because the effect is constant. As an ATE estimator, single-proxy control also fits the bridge on the treated units, which gives the effect on the untreated, and averages the two effects with weights equal to the shares of treated and untreated units. CEVAE is reimplemented in PyTorch with the architecture and default settings of the Pyro reference implementation: a 20-dimensional latent, three hidden layers of 200 units, 30 epochs, batch size 100, and learning rate $10^{-3}$, fitted on the standardised $(X, W)$. In SCM-4 the comparators that use the proxy receive the six numbers of the image reader, except the pixel CNN, which starts from the reader's weights and fine-tunes them on the images.

The kernel single-proxy benchmark uses the generator of Xu & Gretton (2025) with a hidden $U \sim \mathrm{Unif}(-1, 1)$, a binary treatment, $P(A = 1 \mid U) = 0.1 + 0.8\{1 + \mathrm{erf}(U)\}/2$, and the deterministic outcome $Y = \sin(\pi U/2) + A - 0.3$, so $\tau = 1$. PROBE receives the two proxies of that generator as two blocks, $W_1 = e^U + \epsilon$ with $\epsilon \sim N(0, 0.1^2)$ and $W_2$, an MNIST image of the digit $\lfloor 5U + 5 \rfloor$; KSPC receives one of them at a time. Here KSPC is the authors' code without covariates, and Fig. 7 shows both of its estimators, SKPV and SPMMR. With two splits, PROBE retains a split when its mean gap over the four rotations is at most $z$ times its standard error, with $z = \Phi^{-1}(1 - 0.05/2)$. The benchmark has 20 replicates of size $n = 3{,}000$.

Fig. 6 gives the detailed comparison. PROBE answers every one of the 80 original data sets, with mean absolute error between 0.014 and 0.020 and 90th percentile error between 0.033 and 0.046. Its prespecified one-sided paired bound on the excess over the oracle ranges from 0.015 to 0.023. The two ways of using the proxy without an audit both fail: adjustment for $X$ and the entire raw proxy has mean estimate between 0.55 and 0.73 because it absorbs the treatment-only coordinate, and in SCM-4 a convolutional network that reads every pixel and never tests balance has mean estimate 0.72. The paired bound against that pixel network is $-0.25$.

The prespecified-block contrast isolates identification from search, as Thm. 1 requires. Holding out the clean block $W_1$ and balancing it gives 0.998, 0.987, 0.980 and 1.000 in the four settings. Holding out the contaminated block $W_4$ gives 0.660, 0.656, 0.616 and 0.667, and the balance screen retains that split in 79 of the 80 original data sets and 396 of all 400, because $W_4$ really can be balanced. Holding out $W_5$ gives about 0.29 and is rejected in all 80 original data sets and all 400, since no representation balances a block that carries the treatment-only cause. Balance is observable, and on its own it is not enough.

Aggregation repairs the one failure that balance leaves. Thm. 2 and Cor. 1 predict that the valid splits outvote $W_4$ when they form a majority, and they do: averaging over the retained splits is dragged to 0.963, 0.956, 0.945 and 0.959, the plain median of the retained splits already recovers most of that at 1.002, 0.998, 0.991 and 0.997, and the median of the largest agreeing component that PROBE returns reaches 1.005, 1.003, 0.993 and 1.001. The screen has true positive rate between 0.986 and 1.000 and false positive rate between 0.041 and 0.046, every returned component is pure, and every data set returns a unique largest component.




† run after the protocol was frozen

Figure 6: Detailed version of the left of Fig. 3 for the original 20 data sets per setting, with every method on its own row. KSPC (†) was run after the protocol was frozen.

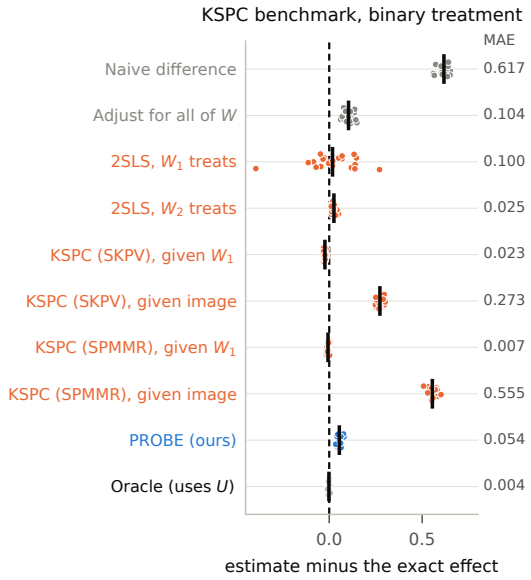



Figure 7: Kernel single-proxy benchmark of Xu & Gretton (2025) with a binary treatment; the effect is exactly 1. Estimates minus the effect over 20 replicates, with the MAE on the right.

The two bounds of Thm. 4 and Thm. 5 are visible separately. For a single split in SCM-1 to SCM-3, the population effect of the learned representation sits a median of 0.042 to 0.046 from $\tau$, with 90th percentile 0.103 to 0.110, and the realised error of that one split has median 0.051 to 0.059. Aggregating the retained splits cuts this to the 0.014 to 0.020 reported above, so the combination step, not a single well-behaved representation, carries the accuracy. The same separation shows up

Table 2: Sealed confirmations. Each row is 20 independently generated data sets at $n = 24{,}000$.

| Setting | What the learner sees | Mean error | p90 | Oracle gap | Status |
|---|---|---|---|---|---|
| SCM-1 | 8 coordinates | 0.020 | 0.033 | 0.022 | 12/12 gates |
| SCM-2 | 88 coordinates | 0.017 | 0.037 | 0.019 | 12/12 gates |
| SCM-3 | 1,312 coordinates | 0.014 | 0.035 | 0.015 | 12/12 gates |
| SCM-4 | six $64 \times 64$ images | 0.018 | 0.046 | 0.023 | 13/13 gates, amended |

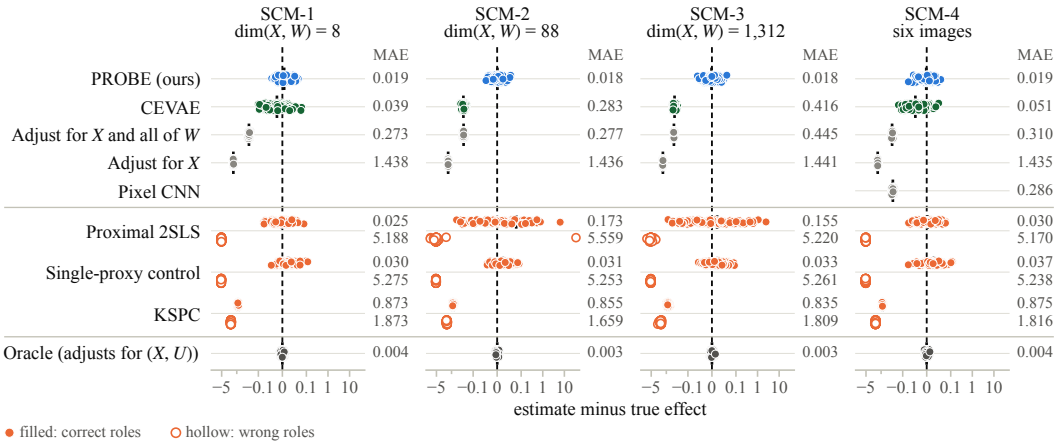



Figure 8: Detailed version of the left of Fig. 3 over all 100 data sets per setting, with the MAE of each method on the right; for methods that need proxy roles, the upper number is for the correct roles.

in the intervals. A nominal 95% interval around one split, built from the estimate's own standard error, covers $\tau_{\widehat{Z}}$ at 0.950, 0.957 and 0.936 in SCM-1 to SCM-3, which is the nominal rate, and covers $\tau$ at 0.836, 0.857, 0.879 and 0.807. The shortfall is the representation gap, which no interval built from sampling variation can contain; Fig. 11 shows it per data set.

Two boundaries limit what these results show. The image setting hands PROBE numbers produced by a reader that was trained with orientation labels on an auxiliary image bank, disjoint from the causal data; that reader made no level error in 14,400,000 reads of all 100 data sets, so SCM-4 tests the pipeline under an image proxy rather than the discovery of a representation from raw pixels. The second boundary is the one the $W_4$ contrast exposes: the screen certifies balance and cannot certify the outcome-relevant condition that Assumption 3 adds.

The SCM-4 row carries a correction that we report rather than hide. Two of its 20 data sets reached the wall-clock cap while three tasks shared one GPU, so the pixel-network comparator recorded no value there and the sealed summary returned FAIL on the completeness gate and on the comparison that depends on it. Those two comparator runs were repeated alone under the sealed configuration and the sealed seeds, every other number was left untouched, and the amended summary passes all 13 gates.

### D.7 Semi-synthetic benchmarks

Twins (Louizos et al., 2017), IHDP (Hill, 2011), and ACIC 2016 (Dorie et al., 2019) keep their covariates and both potential outcomes, so the effect is known exactly. In each, one covariate is hidden as $U$ (gestational age in Twins, the sixth of the 25 IHDP covariates, and $x_{44}$ in ACIC), and the five proxy blocks and the treatment are generated from it by one fixed recipe. The blocks follow the proxy mechanism of SCM-1 with a generated standard normal treatment-only cause $I$, and the treatment follows $P(A = 1 \mid U, I, X) = 0.1 + 0.8\,\Phi(1.5U + 2.0I + w^\top \tilde{X})$, where $U$ is standardised, $\tilde{X}$ is the standardised $X$, and $w \sim N(0, 0.1^2 I)$ is drawn once. Twins has $n = 11{,}984$ and 42 covariates, IHDP $n = 747$ and 24, and ACIC $n = 4{,}802$ and 78. The original runs use Twins

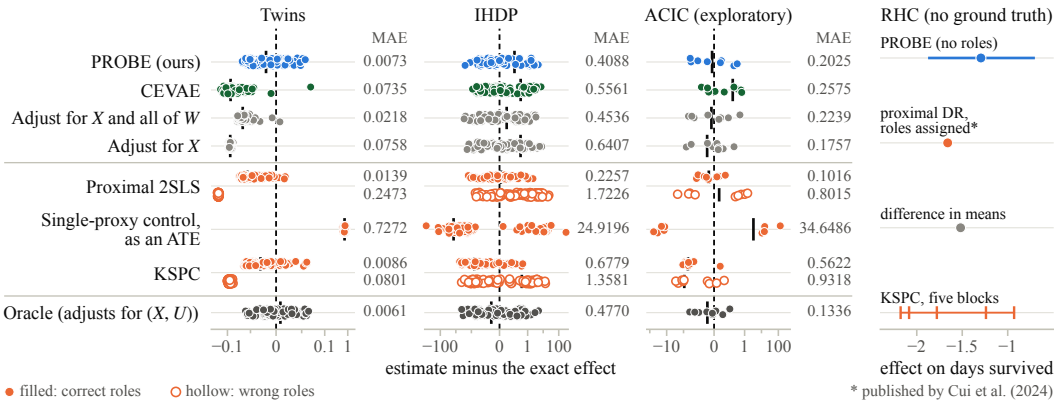



Figure 9: Detailed version of the right of Fig. 3 over 100 Twins, 88 IHDP, and 9 ACIC data sets, with the MAE of each method, and the RHC estimates, where no true effect is known; PROBE's estimate carries its resampled 95% interval, and * marks estimates published by Cui et al. (2024).

replicates 11 to 20, IHDP replicates 12 to 21, and ACIC realizations 2 to 10. The comparators take the proxy roles of Appendix D.6; single-proxy control is run once, as an ATE estimator with the clean block $W_1$.

The sealed thresholds of SCM-1 to SCM-4 do not transfer to these data sets, because the noise level of the screen statistic changes with the numbers of rows and covariates. The screen therefore uses a noise-scaled threshold: instead of requiring a pass in all four rotations, it pools them, taking their mean gap and its standard error, and it retains a split when $\max(\text{gap}, 0) + z\,\text{SE}$ is at most the median of $z\,\text{SE}$ over the thirty splits of the same data set. This rule was adopted after the pre-registered RHC analysis and the RHC-planted, ACIC, and earlier IHDP analyses with the sealed threshold returned no estimate. It was frozen before Twins replicates 11 to 20 and IHDP replicates 12 to 21 were generated, so these two studies are prospective tests; it was chosen after seeing ACIC realizations 2 to 10, so the ACIC results are exploratory.

The RHC study (Connors et al., 1996) has 5,735 patients, days survived within 30 as the outcome, and ten physiological measurements grouped into five proxy blocks, as analyzed by Cui et al. (2024); no true effect is known. The blocks are, in order, the $PaO_2/FiO_2$ ratio and $PaCO_2$, pH and hematocrit, white cell count and albumin, bilirubin and creatinine, and sodium and potassium, and the comparators are the difference in means, the proximal estimate that Cui et al. (2024) obtain by assigning roles to $W_1$ and $W_2$, and KSPC given each block in turn. The interval around PROBE's RHC estimate in Fig. 9 gives the sampling spread of the median of the largest agreeing group over 20,000 draws, holding the retained splits and the linking radius at their observed values, so it omits the uncertainty of the screen.

Table 3 gives the MAE on the original replicates. On IHDP, ten replicates do not resolve the gap between PROBE and CEVAE: the error ratio is 1.67, with a 90% paired-bootstrap interval from 0.78 to 2.78.

### D.8 Protocol discipline

One generated data set is one replicate; splits, rows, rotations, and restarts are not. Seeds overlap across the synthetic settings: SCM-1 replicates 10 to 19 share seeds with SCM-2 replicates 0 to 9, and SCM-2 replicates 10 to 19 with SCM-3 replicates 0 to 9, so comparisons across those settings are not independent. Paired methods share data, sample roles, and nuisance budgets. For SCM-1 to SCM-4 the screen threshold is fixed before the confirmation data are generated, and Appendix D.7 gives the rule and its history for the other data sets. The threshold remains a screen rather than proof of causal validity. No-return outcomes stay in the return-rate denominator, failed sealed runs stay in the record, and this appendix documents exclusions, replacements, and the representation family the search ranges over.

Table 3: Mean absolute error on the original replicates in three semi-synthetic studies with known effects; ACIC is exploratory. Proximal 2SLS, single-proxy control, and KSPC require proxy roles; KSPC (†) was run after the protocol was frozen. Bold marks the smallest error among the methods other than the oracle.

| Method | Twins | IHDP | ACIC |
|---|---|---|---|
| PROBE (ours) | **0.0077** | 0.3054 | 0.2025 |
| CEVAE | 0.0894 | **0.1828** | 0.2575 |
| Adjust for $X$ and all of $W$ | 0.0207 | 0.3927 | 0.2239 |
| Adjust for $X$ | 0.0749 | 0.7221 | 0.1757 |
| Proximal 2SLS, roles given | 0.0156 | 0.2328 | **0.1016** |
| Proximal 2SLS, roles swapped | 0.2453 | 1.5856 | 0.8015 |
| Single-proxy control, as an ATE | 0.7318 | 5.9838 | 34.6486 |
| KSPC$^{\dagger}$, clean block | 0.0099 | 0.6214 | 0.5622 |
| KSPC$^{\dagger}$, contaminated block | 0.0815 | 1.2499 | 0.9318 |
| Oracle (adjusts for $(X, U)$) | 0.0074 | 0.4868 | 0.1336 |

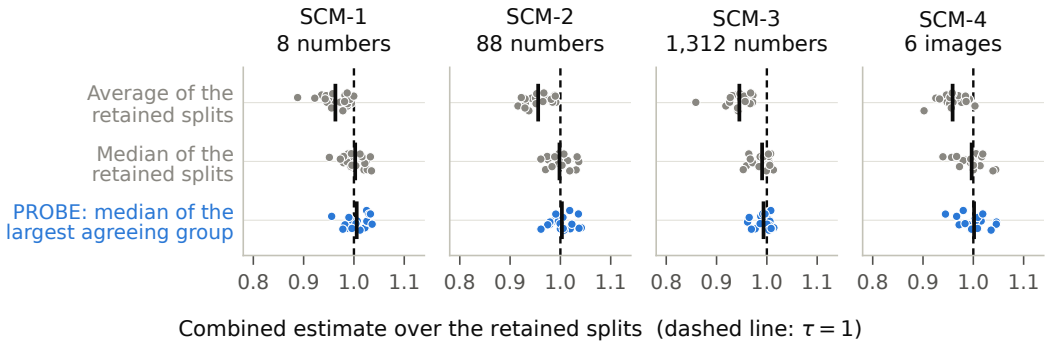



Figure 10: Three ways to combine the retained splits. Averaging is pulled away from $\tau$ by the balanced but invalid split; the median of the largest agreeing group is not.

### D.9 Evidence boundaries

The empirical gates are performance criteria, not proofs of completeness, stability, or the population plurality condition.

### D.10 Additional figures

Fig. 10 shows the three aggregation rules on the sealed confirmations, Fig. 11 shows the interval coverage of one split, and Fig. 12 separates the representation and estimation parts of the error.

The interval in Fig. 11 is the AIPW estimate of one split plus or minus 1.96 times its own standard error, and that standard error is computed from the score cross-product the sealed run already stored, so the figure refits nothing. It covers what the learned representation targets at the nominal rate and covers $\tau$ less often, by the amount of the representation gap. An interval built from sampling variation cannot contain that gap, which is why Thm. 4 bounds it separately from Thm. 5.

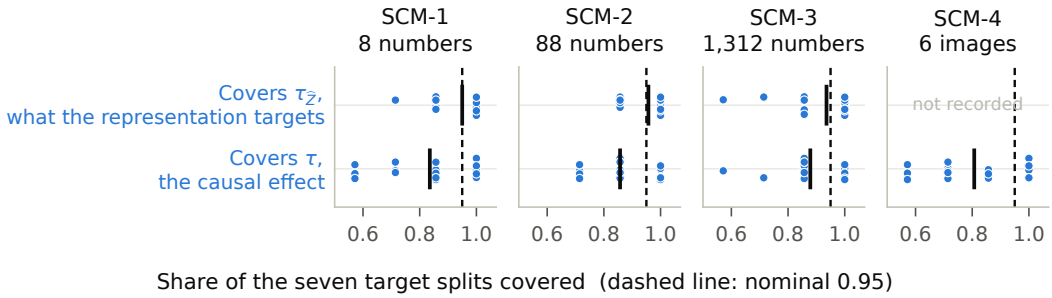



Figure 11: Coverage of a nominal 95% interval around one split, one dot per data set. The interval covers the population effect of the learned representation at the nominal rate and covers $\tau$ less often, by the representation gap.

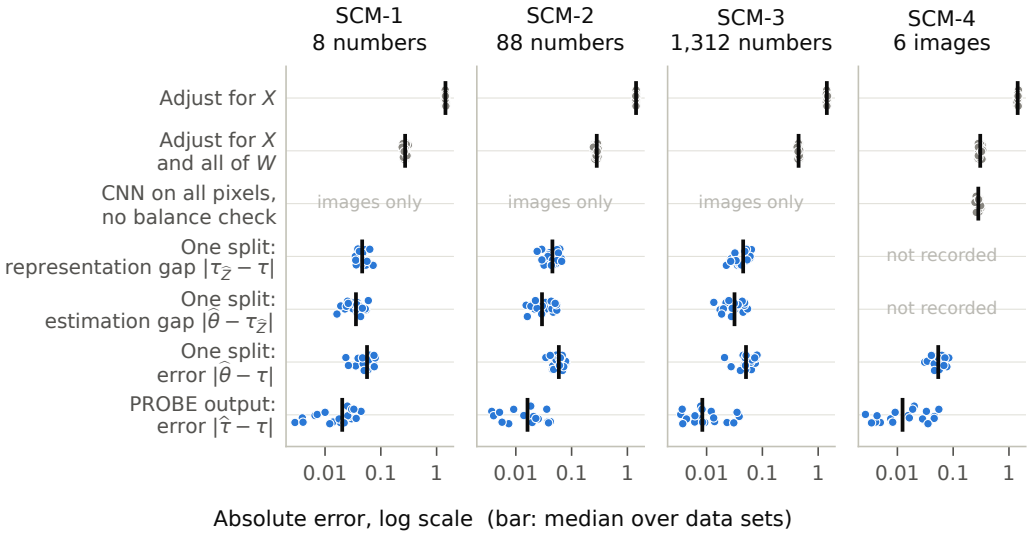



Figure 12: Absolute errors on a log scale. For one split the representation gap $|\tau_{\hat{Z}} - \tau|$ is the population effect of the learned representation, which is exact for the linear representations of SCM-1 to SCM-3 and is not recorded for SCM-4. Aggregation reduces the error of a single split by roughly a factor of three.

# E Proofs

## E.1 Proofs for Section 3

**Sufficient conditions for outcome-relevant completeness.**

*Proof of Proposition 2.* Fix $z$ and $Q, Q' \in \mathcal{Q}_{j,z}$ with $\mathcal{K}_j Q = \mathcal{K}_j Q'$. Write $\pi_Q$ for the distribution of $(X, U)$ under $Q$, $\pi_{Q,X}$ for its $X$-marginal, and $\pi_Q(\cdot \mid x)$ for its conditional distribution of $U$ given $X = x$. By condition (i), $K_j(\cdot \mid v, u) = K_j(\cdot \mid x, u)$, so the distribution $\mathcal{K}_j Q$ of $(X, W_j)$ has $X$-marginal $\pi_{Q,X}$ and conditional distribution $\int K_j(\cdot \mid x, u)\, \pi_Q(du \mid x)$ given $X = x$. The equality $\mathcal{K}_j Q = \mathcal{K}_j Q'$ therefore gives $\pi_{Q,X} = \pi_{Q',X}$ and, for $\pi_{Q,X}$-almost every $x$, $\int K_j(\cdot \mid x, u)\, \pi_Q(du \mid x) = \int K_j(\cdot \mid x, u)\, \pi_{Q'}(du \mid x)$. Condition (ii) yields $\pi_Q(\cdot \mid x) = \pi_{Q'}(\cdot \mid x)$ for almost every $x$, hence $\pi_Q = \pi_{Q'}$. By condition (i), $\mu_{a,j,z}(v, u) = \bar{\mu}_{a,z}(x, u)$ for a measurable function $\bar{\mu}_{a,z}$, so $\mathbb{E}_Q\{\mu_{a,j,z}(V_j, U)\} = \int \bar{\mu}_{a,z}\, d\pi_Q = \int \bar{\mu}_{a,z}\, d\pi_{Q'} = \mathbb{E}_{Q'}\{\mu_{a,j,z}(V_j, U)\}$, which is Assumption 3.

For case (a), the distribution of $W_j$ given $X = x$ is the convolution of the pushforward of $\pi$ under $g(x, \cdot)$ with the distribution of $\epsilon_j$. Its characteristic function is the product of the characteristic

function of the pushforward and that of $\epsilon_j$, and the latter vanishes nowhere by assumption. Two distributions $\pi$ and $\pi'$ with equal convolutions therefore have pushforwards with equal characteristic functions, hence equal pushforwards by the uniqueness theorem for characteristic functions, and injectivity of $g(x,\cdot)$ gives $\pi = \pi'$. Thus (ii) holds.

For case (b), the probability generating function of $W_j$ given $X = x$ under a distribution $\pi$ of $U$ is $\mathbb{E}(\prod_\ell s_\ell^{W_{j,\ell}}) = \int \exp\{-\sum_\ell \lambda_\ell(x,u)(1-s_\ell)\}\,\pi(du)$ for $s \in [0,1]^d$, which is the Laplace transform of the pushforward of $\pi$ under $\lambda(x,\cdot)$ evaluated on $[0,1]^d$. A finite measure on $[0,\infty)^d$ is determined by its Laplace transform on a set with nonempty interior, so two distributions with equal mixtures have equal pushforwards, and injectivity of $\lambda(x,\cdot)$ gives $\pi = \pi'$. In the scalar case this is the identifiability of Poisson mixtures of Teicher (1961). Thus (ii) holds.

□

**Identification from a held-out audit.**

*Proof of Theorem 1.* Because $Z_j = \phi_j(V_j)$ is a measurable function of $V_j$, conditioning on $(U, V_j, Z_j)$ is equivalent to conditioning on $(U, V_j)$. The two clauses of Assumption 2 therefore remain valid after additionally conditioning on $Z_j$. Thus, for each $a \in \{0,1\}$,

$$\begin{aligned} \mathbb{E}\{Y(a) \mid A, V_j, U, Z_j\} &= \mathbb{E}\{Y(a) \mid V_j, U, Z_j\}, \\ P(W_j \mid A, V_j, U, Z_j) &= K_j(\cdot \mid V_j, U). \end{aligned} \tag{33}$$

Integrating the second identity over $Q_{b,j,z}$ and using the definition of $\mathcal{K}_j$ gives $P(X, W_j \mid A = b, Z_j = z) = \mathcal{K}_j Q_{b,j,z}$. Exact balance in Eq. (1) therefore implies $\mathcal{K}_j Q_{1,j,z} = \mathcal{K}_j Q_{0,j,z}$ for $P_{Z_j}$-almost every $z$. Representation overlap gives $Q_{b,j,z} \ll P(V_j, U \mid Z_j = z)$ for $b \in \{0,1\}$. Assumption 1, Eq. (33), and Jensen's inequality give the required integrability, so $Q_{b,j,z} \in \mathcal{Q}_{j,z}$ almost everywhere. Assumption 3 therefore gives, for each $a \in \{0,1\}$,

$$\mathbb{E}_{Q_{1,j,z}}\{\mu_{a,j,z}(V_j, U)\} = \mathbb{E}_{Q_{0,j,z}}\{\mu_{a,j,z}(V_j, U)\}. \tag{34}$$

By Eq. (33) and iterated expectation, for each $b \in \{0,1\}$, $\mathbb{E}\{Y(a) \mid A = b, Z_j = z\} = \mathbb{E}_{Q_{b,j,z}}\{\mu_{a,j,z}(V_j, U)\}$. Eq. (34) shows that the conditional mean of $Y(a)$ given $(A, Z_j)$ does not depend on $A$. Consistency then gives $\mathbb{E}(Y \mid A = a, Z_j = z) = \mathbb{E}\{Y(a) \mid A = a, Z_j = z\} = \mathbb{E}\{Y(a) \mid Z_j = z\}$. Integrating over $P_{Z_j}$ yields Eq. (2). □

**Non-nesting of canonical proximal and proximal-balancing sufficient conditions.**

*Proof of Proposition 3.* We give one finite-state distribution for each direction. In both constructions, $X$ is constant, $Y = Y(A)$, and every treatment draw is generated by an exogenous randomizer independent of all proxy noises and outcome noises.

First, we construct a distribution satisfying the canonical proximal conditions but not the proximal-balancing conditions for the designated proxy split. Let $U \sim \mathrm{Bernoulli}(1/2)$, $P(A = 1 \mid U = u) = \frac{1}{4} + \frac{1}{2}u$, and let $Z^{\mathrm{tr}} = U \oplus E_Z, \qquad Z^{\mathrm{out}} = W_j = U \oplus E_W$, where $E_Z$ and $E_W$ are mutually independent $\mathrm{Bernoulli}(1/4)$ variables, independent of the treatment randomizer. Set $Y(a) = a + U$ and $V_j = Z^{\mathrm{tr}}$. Consistency, integrability, and strict latent overlap hold. Independence of the exogenous noises gives the three canonical proxy-validity conditions.

The function $h(z^{\mathrm{out}}, a) = a + 2z^{\mathrm{out}} - \frac{1}{2}$ is an outcome bridge because $\mathbb{E}(Z^{\mathrm{out}} \mid U = u) = 1/4 + u/2$, so $\mathbb{E}\{h(Z^{\mathrm{out}}, a) \mid U = u\} = a + u$. Iterated expectation gives the required observed bridge equation. The binary channel from $U$ to $Z^{\mathrm{tr}}$ has matrix $\begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix}$ and determinant $1/2$. Conditional on either treatment arm, both latent states have positive probability, and the corresponding unnormalized conditional-expectation matrix has determinant $3/32$. It therefore has full column rank, which proves the canonical treatment-proxy completeness condition.

Now consider any measurable representation $Z_j = \phi_j(V_j)$. The Markov relation $U \to V_j \to Z_j$ holds. Conditional on $(U, Z_j)$, treatment and $W_j$ are independent, and $\mathbb{E}(A \mid U, Z_j) = \mathbb{E}(W_j \mid U, Z_j) = \frac{1}{4} + \frac{1}{2}U$. The conditional covariance decomposition gives $\mathrm{Cov}(A, W_j \mid Z_j) = \frac{1}{4}\mathrm{Var}(U \mid Z_j)$. Because $P(U = 1 \mid V_j = 0) = 1/4$ and $P(U = 1 \mid V_j = 1) = 3/4$, $P(U = 1 \mid$

$Z_j) = \mathbb{E}\{P(U = 1 \mid V_j) \mid Z_j\} \in [1/4, 3/4]$ almost surely. Hence $\mathrm{Var}(U \mid Z_j) \geq 3/16$ and $\mathrm{Cov}(A, W_j \mid Z_j) \geq \frac{3}{64} > 0$. Exact balance therefore fails for every such representation. This proves the first direction.

Second, we construct a distribution satisfying the proximal-balancing conditions for the designated proxy split but not the canonical proximal conditions. Let $U = (U_1, U_2)$, where $U_1$ and $U_2$ are independent $\mathrm{Bernoulli}(1/2)$ variables. Let $Z^{\mathrm{tr}}$ and $Z^{\mathrm{out}}$ be separate observed coordinates that agree almost surely with $U_1$, and set $V_j = Z^{\mathrm{tr}} = U_1, \qquad W_j = Z^{\mathrm{out}} = U_1, \qquad Z_j = U_1$. Define $P(A = 1 \mid U_1, U_2) = \frac{3}{20} + \frac{1}{5}U_1 + \frac{2}{5}U_2$, and let $\xi$ be an independent mean-zero Rademacher variable, also independent of the treatment randomizer. Set $Y(a) = a + U_1 + (1 + U_2)\xi$. The four latent treatment probabilities are $3/20$, $11/20$, $7/20$, and $15/20$, so strict latent overlap holds. Latent exchangeability follows from the independent outcome noise, and held-out treatment independence holds because $W_j$ is deterministic given $U$. Moreover, $P(A = 1 \mid Z_j = 0) = \frac{7}{20}, \qquad P(A = 1 \mid Z_j = 1) = \frac{11}{20}$, so representation overlap holds. Because $W_j = Z_j$ and $X$ is constant, exact balance also holds.

It remains to verify the all-$Q$ outcome-relevant completeness condition. Within a stratum $Z_j = z$, every admissible distribution $Q \in \mathcal{Q}_{j,z}$ is supported on $V_j = U_1 = z$, while $U_2$ may vary. On this support, $\mu_{a,j,z}(v, u) = a + z$. Thus, for any admissible $Q, Q'$ and either potential treatment level, $\mathcal{K}_j Q = \mathcal{K}_j Q' \implies \mathbb{E}_Q\{\mu_{a,j,z}(V_j, U)\} = a + z = \mathbb{E}_{Q'}\{\mu_{a,j,z}(V_j, U)\}$. This proves all proximal-balancing conditions for the designated proxy split.

The outcome bridge $h(z^{\mathrm{out}}, a) = a + z^{\mathrm{out}}$ exists because $\mathbb{E}(Y \mid Z^{\mathrm{tr}} = z, A = a) = a + z = \mathbb{E}\{h(Z^{\mathrm{out}}, a) \mid Z^{\mathrm{tr}} = z, A = a\}$. Canonical treatment-proxy completeness nevertheless fails. In the treated arm, $P(U_2 = 1 \mid U_1 = 0, A = 1) = \frac{11}{14}, \qquad P(U_2 = 1 \mid U_1 = 1, A = 1) = \frac{15}{22}$. Define the square-integrable, nonzero function $g_1(U) = U_2 - \begin{cases} 11/14, & U_1 = 0, \\ 15/22, & U_1 = 1. \end{cases}$ Then $\mathbb{E}\{g_1(U) \mid Z^{\mathrm{tr}}, A = 1\} = \mathbb{E}\{g_1(U) \mid U_1, A = 1\} = 0$ almost surely, although $g_1(U)$ is not almost surely zero. Hence the canonical completeness condition fails, proving the second direction.

For the final claim, the second construction satisfies Assumption 3 but not canonical treatment-proxy completeness. In the first construction, replace $E_W$ by an independent $\mathrm{Bernoulli}(1/2)$ variable. Canonical treatment-proxy completeness is unchanged because it involves only $(U, Z^{\mathrm{tr}}, A)$. Now $W_j$ is independent of $U$, so $\mathcal{K}_j Q$ is the same distribution for every $Q$. For every representation $\phi_j$ and almost every $z$, the bound $P(U = 1 \mid Z_j) \in [1/4, 3/4]$ shown above gives both values of $U$ positive probability given $Z_j = z$, so $\mathbb{E}_Q \mu_{a,j,z} = a + Q(U = 1)$ takes more than one value over $\mathcal{Q}_{j,z}$ and Assumption 3 fails. □

**Identification with an unknown valid block.**

*Proof of Theorem 2.* Under condition (4), the number of proxy splits in $\mathcal{B}$ whose candidate value equals $\tau$ is strictly larger than the count for every non-target value. Therefore, $\tau$ is the unique value maximizing the population count: $\{\tau\} = \arg\max_{c \in \mathbb{R}} \big|\{S \in \mathcal{B} : \theta_S = c\}\big|$. If no non-target value occurs, the convention in Definition 2 compares the nonempty target cluster with zero and gives the same conclusion. □

**Target-valued majority.**

*Proof of Corollary 1.* Define $N_\tau \triangleq \big|\{S \in \mathcal{B} : \theta_S = \tau\}\big|$ and $N_c \triangleq \big|\{S \in \mathcal{B} : \theta_S = c\}\big|, \qquad c \neq \tau$. Condition (17) gives $N_\tau > |\mathcal{B}|/2$. For every $c \neq \tau$, $N_c \leq |\mathcal{B}| - N_\tau < N_\tau$. Therefore, $\tau$ is the unique candidate value of greatest multiplicity.

It remains to establish the median statement. Let $n = |\mathcal{B}|$. If $n = 2k+1$ is odd, then $N_\tau \geq k+1$. At most $k$ candidate values are strictly below $\tau$, and at most $k$ are strictly above $\tau$, so the $(k+1)$st order statistic equals $\tau$. If $n = 2k$ is even, then $N_\tau \geq k+1$ and $n - N_\tau \leq k-1$. Hence both the $k$th and the $(k+1)$st order statistics equal $\tau$. Under the conventional definition of the even-sample median as the average of these two middle order statistics, the median equals $\tau$. Thus, $\mathrm{median}\big((\theta_S)_{S \in \mathcal{B}}\big) = \tau$. The stronger condition that more than half of the proxy splits in $\mathcal{B}$ satisfy the grouped assumptions is sufficient because each such proxy split has candidate value $\tau$, but it is not necessary. □

**Zero discrepancy.**

*Proof of Proposition 1.* By Definition 3, $D_{\mathrm{res},j}(\phi) = 0$ if and only if $q_{j,\phi}(T_j, Z_j^\phi) = e_{j,\phi}(Z_j^\phi)$ almost surely. Because $A$ is binary, this equality is equivalent to $T_j \perp\!\!\!\perp A \mid Z_j^\phi$. □

**Supporting bias decomposition.** Fix a representation $\phi$ and abbreviate $e(z) = e_{j,\phi}(z)$. For each $a \in \{0,1\}$ and $P_{Z_j^\phi}$-almost every $z$, define $\delta_{a,j}^\phi(z) \triangleq \mathbb{E}_{Q_{1,j,z}^\phi}\{\mu_{a,j,z}^\phi(V_j,U)\} - \mathbb{E}_{Q_{0,j,z}^\phi}\{\mu_{a,j,z}^\phi(V_j,U)\}$. Under Assumptions 1 and 2, and under $0 < e(Z_j^\phi) < 1$ almost surely, the conditioning argument used in Eq. (33) gives $\mathbb{E}(Y \mid A=a, Z_j^\phi = z) = \mathbb{E}_{Q_{a,j,z}^\phi}\{\mu_{a,j,z}^\phi(V_j,U)\}$. Conditional expectation over $A$ gives $\mathbb{E}\{Y(a) \mid Z_j^\phi = z\} = e(z)\mathbb{E}_{Q_{1,j,z}^\phi}\{\mu_{a,j,z}^\phi\} + \{1-e(z)\}\mathbb{E}_{Q_{0,j,z}^\phi}\{\mu_{a,j,z}^\phi\}$. Subtracting the two conditional expressions and collecting terms gives

$$\begin{aligned}&\mathbb{E}(Y \mid A=1, Z_j^\phi = z) - \mathbb{E}(Y \mid A=0, Z_j^\phi = z)\\ &\quad - \Big[\mathbb{E}\{Y(1) \mid Z_j^\phi = z\} - \mathbb{E}\{Y(0) \mid Z_j^\phi = z\}\Big]\\ &\quad = \{1-e(z)\}\delta_{1,j}^\phi(z) + e(z)\delta_{0,j}^\phi(z). \end{aligned} \tag{35}$$

**Approximate-balance bias.**

*Proof of Theorem 3.* Fix $z$ in the almost-sure set on which overlap and the structural conditions hold. Bayes' rule gives $\frac{dP(T_j \mid A=1, Z_j^\phi = z)}{dP_{\phi,z}}(t) = \frac{q_{j,\phi}(t,z)}{e(z)}$ and $\frac{dP(T_j \mid A=0, Z_j^\phi = z)}{dP_{\phi,z}}(t) = \frac{1-q_{j,\phi}(t,z)}{1-e(z)}$. Subtracting gives the density identity $\frac{d\{P(T_j \mid A=1, Z_j^\phi = z) - P(T_j \mid A=0, Z_j^\phi = z)\}}{dP_{\phi,z}}(t) = \frac{q_{j,\phi}(t,z) - e(z)}{e(z)\{1-e(z)\}}$. Held-out treatment independence makes the conditional distribution of $W_j$ given $(U, V_j)$ treatment-invariant, so $\mathcal{K}_j Q_{b,j,z}^\phi = P(T_j \mid A = b, Z_j^\phi = z)$ for $b \in \{0,1\}$. As in the proof of Thm. 1, overlap and Assumption 1 give $Q_{b,j,z}^\phi \in \mathcal{Q}_{j,z}^\phi$ for $b \in \{0,1\}$. Assumption 4 therefore implies

$$|\delta_{a,j}^\phi(z)| \le \frac{\Gamma_j(\phi)}{e(z)\{1-e(z)\}}\left[\mathbb{E}\Big(\{q_{j,\phi}(T_j,z) - e(z)\}^2 \Big| Z_j^\phi = z\Big)\right]^{1/2}. \tag{36}$$

Apply the triangle inequality to Eq. (35), use Eq. (36) for both potential outcomes, and integrate over $Z_j^\phi$. This gives $\Big|\mathbb{E}\Big[\mathbb{E}(Y \mid A=1, Z_j^\phi) - \mathbb{E}(Y \mid A=0, Z_j^\phi)\Big] - \tau\Big| \le \mathbb{E}\left[\frac{\Gamma_j(\phi)}{e_{j,\phi}(Z_j^\phi)\{1-e_{j,\phi}(Z_j^\phi)\}}\left\{\mathbb{E}\Big[\{q_{j,\phi}(T_j,Z_j^\phi) - e_{j,\phi}(Z_j^\phi)\}^2 \Big| Z_j^\phi\Big]\right\}^{1/2}\right] \le \frac{\Gamma_j(\phi)}{\eta(1-\eta)} D_{\mathrm{res},j}(\phi)$. Because $\Gamma_j(\phi)$ does not depend on $z$, the overlap bound $\eta \le e_{j,\phi}(z) \le 1-\eta$ gives the last inequality, together with $\mathbb{E}\{s(Z_j^\phi)\} \le [\mathbb{E}\{s(Z_j^\phi)^2\}]^{1/2} = D_{\mathrm{res},j}(\phi)$ by Cauchy–Schwarz, where $s(z)$ denotes the conditional root-mean-square factor in the display, and proves Eq. (6). □

**Population sensitivity region.**

*Proof of Proposition 5.* Thm. 3 gives $|\tau_{Z_j^\phi} - \tau| \le \Gamma_j(\phi) D_{\mathrm{res},j}(\phi)/\{\eta(1-\eta)\}$, and the displayed interval is the set of values of $\tau$ compatible with this inequality. □

### E.2 Proofs for Section 4

**Uniform discrepancy deviation.**

*Proof of Proposition 6.* For clarity, the joint encoder–critic loss classes used in the proposition are $\mathcal{L}_{0,j,n_{\mathrm{D}}} \triangleq \big\{o \mapsto \{a - h(\phi(v_j))\}^2 : \phi \in \mathcal{F}_{j,n_{\mathrm{D}}},\ h \in \mathcal{H}_{0,n_{\mathrm{D}}}\big\}$; $\mathcal{L}_{1,j,n_{\mathrm{D}}} \triangleq \big\{o \mapsto \{a - h(t_j, \phi(v_j))\}^2 : \phi \in \mathcal{F}_{j,n_{\mathrm{D}}},\ h \in \mathcal{H}_{1,n_{\mathrm{D}}}\big\}$, where $o = (v_j, t_j, a, y)$. For a loss class $\mathcal{L}$, its expected Rademacher complexity is $\mathfrak{R}_{n_{\mathrm{D}}}(\mathcal{L}) \triangleq \mathbb{E}\Big[\sup_{\ell \in \mathcal{L}} \frac{1}{n_{\mathrm{D}}} \sum_{i=1}^{n_{\mathrm{D}}} \sigma_i \ell(O_i)\Big]$, where

the $\sigma_i$ are independent Rademacher signs independent of the observations. The loss classes are pointwise measurable and separable, so these suprema are measurable. The same argument may instead be stated with outer probability and outer expectation.

Every loss in either class lies in $[0,1]$. Standard symmetrization and bounded-difference concentration therefore imply, simultaneously for $b \in \{0,1\}$, that

$$\sup_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}},\, h \in \mathcal{H}_{b,n_{\mathrm{D}}}} |\widehat{R}_{b,\phi}(h) - R_{b,\phi}(h)| \le 2\mathfrak{R}_{n_{\mathrm{D}}}(\mathcal{L}_{b,j,n_{\mathrm{D}}}) + \sqrt{\frac{\log(4/\delta_{\mathrm{D}})}{2n_{\mathrm{D}}}} \tag{37}$$

with probability at least $1-\delta_{\mathrm{D}}$. Let $R^\star_{b,\phi} = \inf_{h \in \mathcal{H}_{b,n_{\mathrm{D}}}} R_{b,\phi}(h)$ and define $\widehat{R}^\star_{b,\phi}$ analogously. Eq. (37) controls $|\widehat{R}^\star_{b,\phi} - R^\star_{b,\phi}|$ by the same right-hand side. Eq. (18) also gives $0 \le \widehat{R}_{b,\phi}(\widehat{h}_{b,\phi}) - \widehat{R}^\star_{b,\phi} \le \varepsilon_{b,\phi}$.

The conditional-mean property of squared loss yields the Brier identity $\mathbb{E}[\{A - e_{j,\phi}(Z_j^\phi)\}^2] - \mathbb{E}[\{A - q_{j,\phi}(T_j, Z_j^\phi)\}^2] = \mathbb{E}[\{q_{j,\phi}(T_j, Z_j^\phi) - e_{j,\phi}(Z_j^\phi)\}^2] = D^2_{\mathrm{res},j}(\phi)$. It follows from Eq. (19) that

$$\left|R^\star_{0,\phi} - R^\star_{1,\phi} - D^2_{\mathrm{res},j}(\phi)\right| \le a_{0,n_{\mathrm{D}}}(\phi) + a_{1,n_{\mathrm{D}}}(\phi) \le a_{j,n_{\mathrm{D}}}. \tag{38}$$

Combine Eqs. (37)–(38). The positive-part map is one-Lipschitz and $D^2_{\mathrm{res},j}(\phi) \ge 0$, so taking the supremum over $\phi$ gives Eq. (20). □

**Representation-learning oracle inequality.**

*Proof of Proposition 7.* On the event in Eq. (20), set $\bar{\xi}_{j,n_{\mathrm{D}}} \triangleq r_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) + a_{j,n_{\mathrm{D}}} + \varepsilon_{\mathrm{crit},j}$. The deviation inequality at $\widehat{\phi}_j$, Eq. (21), and the deviation inequality at a near-minimizer of $D^2_{\mathrm{res},j}$ over $\mathcal{F}_{j,n_{\mathrm{D}}}$ give $D^2_{\mathrm{res},j}(\widehat{\phi}_j) \le \widetilde{D}^2_{j,n_{\mathrm{D}}}(\widehat{\phi}_j) + \bar{\xi}_{j,n_{\mathrm{D}}} \le \inf_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} \widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) + \varepsilon_{\mathrm{enc},j} + \bar{\xi}_{j,n_{\mathrm{D}}} \le \inf_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) + 2\bar{\xi}_{j,n_{\mathrm{D}}} + \varepsilon_{\mathrm{enc},j}$, which proves Eq. (22). □

**Learned-representation error.**

*Proof of Theorem 4.* Conditional on the training observations $\{O_i : i \in \mathcal{I}_{\mathrm{D}}\}$ and any algorithmic seed used in training, the selected encoder is fixed; the population expectations below use a fresh observation independent of these training inputs. Eq. (6) gives $|\tau_{\widehat{Z}_j} - \tau| \le \frac{\Gamma_j(\widehat{\phi}_j)}{\eta(1-\eta)} D_{\mathrm{res},j}(\widehat{\phi}_j)$. On the event in Eq. (8), which has probability at least $1-\delta_{\mathrm{D}}$, every $\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}$ satisfies $D^2_{\mathrm{res},j}(\widehat{\phi}_j) \le \widetilde{D}^2_{j,n_{\mathrm{D}}}(\widehat{\phi}_j) + \xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) \le \widetilde{D}^2_{j,n_{\mathrm{D}}}(\phi) + \xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}}) \le D^2_{\mathrm{res},j}(\phi) + 2\xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})$. The first and last inequalities hold on this event, and the middle one holds because $\widehat{\phi}_j$ minimizes $\widetilde{D}^2_{j,n_{\mathrm{D}}}$ over $\mathcal{F}_{j,n_{\mathrm{D}}}$. Taking the infimum over $\phi$ and then the square root gives $D_{\mathrm{res},j}(\widehat{\phi}_j) \le [\inf_{\phi \in \mathcal{F}_{j,n_{\mathrm{D}}}} D^2_{\mathrm{res},j}(\phi) + 2\xi_{j,n_{\mathrm{D}}}(\delta_{\mathrm{D}})]^{1/2}$. Substituting this bound into the first display proves Eq. (9). Substituting the square root of Eq. (22) instead proves Thm. 10. □

**Exact signed nuisance remainder.**

*Proof of Proposition 8.* Fix $\widehat{Z}_j = z$. Taking the conditional expectation of the score and using Eq. (24) gives $\mathbb{E}\{\mathtt{UIF}_j(O) \mid \widehat{Z}_j = z\} = \widehat{m}_{1,j}(z) - \widehat{m}_{0,j}(z) + \frac{e_j(z)\{m_{1,j}(z) - \widehat{m}_{1,j}(z)\}}{\widehat{e}_j(z)} - \frac{\{1-e_j(z)\}\{m_{0,j}(z) - \widehat{m}_{0,j}(z)\}}{1-\widehat{e}_j(z)}$. Subtract $m_{1,j}(z) - m_{0,j}(z)$ and collect the two nuisance errors. Averaging the resulting identity over $\widehat{Z}_j$ proves Eq. (25). □

**Honest AIPW error.**

*Proof of Theorem 5.* Conditional on the two learning samples, the evaluation observations are independent and the score $\mathtt{UIF}_j(O)$ has variance $\sigma_j^2$, so Chebyshev's inequality bounds the difference between the evaluation average and its conditional expectation by $\sigma_j/\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E}}}$, the first term in

Eq. (26), with probability at least $1-\delta_{\mathrm{E}}$. Prop. 8, Cauchy–Schwarz, and the propensity clipping bound control the absolute conditional remainder by $\frac{q_{e,j}(q_{0,j}+q_{1,j})}{\eta}$. The triangle inequality proves Eq. (27). For Eq. (13), the triangle inequality gives $|\widehat{\tau}_j^{\mathrm{AIPW}} - \tau| \le |\widehat{\tau}_j^{\mathrm{AIPW}} - \tau_{\widehat{Z}_j}| + |\tau_{\widehat{Z}_j} - \tau|$. Eq. (27) bounds the first term on the same event, and Eq. (6) bounds the second term by $\Gamma_j(\widehat{\phi}_j) D_{\mathrm{res},j}(\widehat{\phi}_j)/\{\eta(1-\eta)\}$. □

**Consistency.**

*Proof of Corollary 2.* Condition (C1) makes the evaluation term $\sigma_{j,n}/\sqrt{n_{\mathrm{E}}\delta_{\mathrm{E},n}}$ of $\varepsilon_{\mathrm{AIPW},j}$ vanish, condition (C2) makes its nuisance-product term vanish, and condition (C3) makes the stability-weighted residual-discrepancy term in Eq. (13) vanish. The failure probability $\delta_{\mathrm{E},n}$ of Thm. 5 also converges to zero, which proves convergence in probability. □

**Cluster plurality under random subsampling.**

*Proof of Lemma 1.* Conditional on $\mathcal{G}$ and $R = r \ge 1$, the retained proxy splits form a uniformly sampled $r$-element subset of $\widehat{\mathcal{B}}$ because the $m$ draws are uniform without replacement and $\widehat{\mathcal{B}}$ is fixed. For each competing cell $C_\ell$, define $H_\ell(S) \triangleq \mathbf{1}\{S \in C_0\} - \mathbf{1}\{S \in C_\ell\}$ for $S \in \widehat{\mathcal{B}}$. This variable lies in $[-1,1]$ and has population mean $(n_0 - n_\ell)/N \ge \Delta_\Pi$. Let $\overline{H}_{\ell,r}$ be its retained-sample average. The event $R_\ell \ge R_0$ implies $\overline{H}_{\ell,r} \le 0$. Hoeffding's inequality for sampling without replacement (Hoeffding, 1963) gives $P(\overline{H}_{\ell,r} \le 0 \mid \mathcal{G}, R = r) \le e^{-r\Delta_\Pi^2/2}$. A union bound over the $L$ competing cells yields $P(\max_{1\le\ell\le L} R_\ell \ge R_0 \mid \mathcal{G}, R = r) \le L e^{-r\Delta_\Pi^2/2}$. Averaging over $R$ and counting $R = 0$ as failure gives $P(\text{failure} \mid \mathcal{G}) \le P(R = 0 \mid \mathcal{G}) + L\,\mathbb{E}\{e^{-R\Delta_\Pi^2/2} \mid \mathcal{G}\}$. If $L = 0$, every retained split belongs to $C_0$, so the conclusion holds whenever $R \ge 1$. □

**Finite-sample PROBE cluster recovery.**

*Proof of Theorem 6.* Work conditionally on $\mathcal{G}$. Let $E_1$ be the uniform-evaluation-error event and let $E_2$ be the event that $R \ge 1$ and $R_0 > \max_{1\le\ell\le L} R_\ell$. Eq. (27) applied to each $S \in \widehat{\mathcal{B}}$ at $\delta_{\mathrm{E}} = \delta/N$ gives $P(|\widehat{\theta}_S - \theta_S| > \varepsilon \mid \mathcal{G}) \le \delta/N$, because $\varepsilon$ is at least the radius of Eq. (26) for $S$ at that level, and a union bound over the $N$ splits in $\widehat{\mathcal{B}}$ gives $P(E_1^{\mathrm{c}} \mid \mathcal{G}) \le \delta$, while Lemma 1 gives $P(E_2^{\mathrm{c}} \mid \mathcal{G}) \le P(R = 0 \mid \mathcal{G}) + L\,\mathbb{E}\{e^{-R\Delta_\Pi^2/2} \mid \mathcal{G}\}$. For $S, S' \in C_0$, Eq. (6) for $S$ with encoder $\widehat{\phi}_S$ and Assumption 5 give $|\theta_S - \tau| \le \mathtt{bias}_S \le b$, and likewise $|\theta_{S'} - \tau| \le b$. Hence $|\theta_S - \theta_{S'}| \le 2b \le 2(\rho - \varepsilon)$ by condition (i), so the within-cluster bound of condition (ii) also holds on $C_0$. On $E_1$, two retained splits in the same cell satisfy $|\widehat{\theta}_S - \widehat{\theta}_{S'}| \le 2(\rho-\varepsilon) + 2\varepsilon = 2\rho$, so they are linked. Two retained splits in different cells satisfy $|\widehat{\theta}_S - \widehat{\theta}_{S'}| > 2(\rho+\varepsilon) - 2\varepsilon = 2\rho$, so they are not linked. Hence the connected components are exactly the nonempty sets $\mathcal{R}_m \cap C_\ell$. On $E_2$, $\mathcal{R}_m \cap C_0$ is the unique largest component, and Algorithm 1 returns the median of its estimates. For every $S \in C_0$, $|\widehat{\theta}_S - \tau| \le |\widehat{\theta}_S - \theta_S| + |\theta_S - \tau| \le \varepsilon + b$, where $|\theta_S - \tau| \le b$ by Eq. (6), because every $S \in C_0$ satisfies Assumptions 1, 2, and 4. The median therefore lies in $[\tau - (b+\varepsilon), \tau + (b+\varepsilon)]$. A union bound over $E_1^{\mathrm{c}}$ and $E_2^{\mathrm{c}}$ proves Eq. (14). □

*Proof of Corollary 3.* Write $c \triangleq \Delta_\Pi^2/2$. First, $P(R = 0) = \prod_{i=0}^{m-1}(T - N - i)/(T - i) \le (1-p)^m \le e^{-mp}$. Second, $R$ is the sum of the indicators $\mathbf{1}\{S \in \widehat{\mathcal{B}}\}$ over the $m$ splits drawn without replacement. Because $x \mapsto e^{-cx}$ is continuous and convex, Thm. 4 of Hoeffding (1963) bounds $\mathbb{E}\{e^{-cR}\}$ by the corresponding expectation under sampling with replacement. Thus $\mathbb{E}\{e^{-cR}\} \le \{1 - p(1-e^{-c})\}^m \le \exp\{-mp(1-e^{-c})\}$. Since $1 - e^{-c} \le 1$, adding the bound for $P(R = 0)$ gives Eq. (32). Its right side is at most $\alpha$ when $m \ge \log\{(L+1)/\alpha\}/\{p(1-e^{-c})\}$, and Thm. 12 gives $P(|\widehat{\tau} - \tau| \le b + \varepsilon) \ge 1 - \delta - \alpha$. Finally, $1 - e^{-c} \ge c/(1+c) \ge \Delta_\Pi^2/3$ because $0 < \Delta_\Pi \le 1$. If $L = 0$, only the empty-draw bound is needed, and $m \ge \log(1/\alpha)/p$ suffices. □

**Median under majority.**

*Proof of Corollary 4.* On the target-band and uniform-evaluation-error event, every retained split in $C_0$ satisfies $|\widehat{\theta}_S - \tau| \le b + \varepsilon$. If $|\mathcal{R}_m \cap C_0| > R/2$, more than half of all retained estimates lie in $[\tau - (b+\varepsilon), \tau + (b+\varepsilon)]$. Both middle order statistics, and hence their midpoint when needed, lie in this interval. Therefore the median lies within $b + \varepsilon$ of $\tau$. □

### E.3 Finite-Strata MMD Specialization

This subsection records the earlier finite-strata analysis as an independent specialization. It is not used by the continuous-representation results in the main text.

Fix a finite library $\mathcal{F}_j^{\mathrm{fs}} \subseteq \Phi_j$ with $N_j^{\mathrm{fs}} \triangleq |\mathcal{F}_j^{\mathrm{fs}}|$. Every $\phi \in \mathcal{F}_j^{\mathrm{fs}}$ is deterministic and takes values in a common finite set $\mathcal{S}_j$ of size $M_j$. Write $Z_j^\phi = \phi(V_j)$, let $\mathcal{R}_j(\phi)$ be its positive-mass strata, and set $p_{j,z}(\phi) = P(Z_j^\phi = z)$. Assume, uniformly over the library and $z \in \mathcal{R}_j(\phi)$, $P(Z_j^\phi = z) \ge p_{\min} > 0, \eta \le P(A = 1 \mid Z_j^\phi = z) \le 1 - \eta$ for some $\eta \in (0, 1/2]$. Let $\kappa_j$ be a characteristic kernel on $(X, W_j)$ with reproducing kernel Hilbert space $\mathcal{H}_j$ and $\sup_v \sqrt{\kappa_j(v,v)} \le \rho_j < \infty$. For $z \in \mathcal{R}_j(\phi)$, define $\Delta_{j,z}^{\mathrm{MMD}}(\phi) \triangleq \mathrm{MMD}_{\kappa_j}\left\{P(X, W_j \mid A = 1, Z_j^\phi = z), P(X, W_j \mid A = 0, Z_j^\phi = z)\right\}$; $D_j^{\mathrm{MMD}}(\phi) \triangleq \sum_{z \in \mathcal{R}_j(\phi)} p_{j,z}(\phi) \Delta_{j,z}^{\mathrm{MMD}}(\phi)$.

For $i \in \mathcal{I}_{\mathrm{D}}$, let $Z_{j,i}^\phi = \phi(V_{j,i})$ and define $N_{a,z}^\phi \triangleq \sum_{i \in \mathcal{I}_{\mathrm{D}}} \mathbf{1}\{A_i = a, Z_{j,i}^\phi = z\}$; $N_z^\phi \triangleq N_{0,z}^\phi + N_{1,z}^\phi$; $\widehat{p}_{j,z}(\phi) \triangleq \frac{N_z^\phi}{n_{\mathrm{D}}}$; $\widehat{m}_{a,j,z}(\phi) \triangleq \frac{1}{N_{a,z}^\phi} \sum_{i \in \mathcal{I}_{\mathrm{D}}} \mathbf{1}\{A_i = a, Z_{j,i}^\phi = z\}$; $\times \kappa_j\{(X_i, W_{j,i}), \cdot\}$. When both treatment-arm cells are nonempty, set $\widehat{\Delta}_{j,z}^{\mathrm{MMD}}(\phi) \triangleq \|\widehat{m}_{1,j,z}(\phi) - \widehat{m}_{0,j,z}(\phi)\|_{\mathcal{H}_j}$; $\widehat{D}_j^{\mathrm{MMD}}(\phi) \triangleq \sum_{z : N_z^\phi > 0} \widehat{p}_{j,z}(\phi) \widehat{\Delta}_{j,z}^{\mathrm{MMD}}(\phi)$. If an occupied stratum contains only one treatment arm, set $\widehat{D}_j^{\mathrm{MMD}}(\phi) = +\infty$.

The MMD specialization uses its own structural stability condition. For every $\phi \in \mathcal{F}_j^{\mathrm{fs}}$, every positive-mass stratum $z$, every $a \in \{0,1\}$, and all $Q, Q' \in \mathcal{Q}_{j,z}$, define $\Delta_{a,j,z}^{\mu}(Q, Q') \triangleq \mathbb{E}_Q\{\mu_{a,j,z}\} - \mathbb{E}_{Q'}\{\mu_{a,j,z}\}$. Assume

$$\left|\Delta_{a,j,z}^{\mu}(Q, Q')\right| \le \Gamma_j^{\mathrm{MMD}}(\phi)\, \mathrm{MMD}_{\kappa_j}(\mathcal{K}_j Q, \mathcal{K}_j Q'), \qquad \sup_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} \Gamma_j^{\mathrm{MMD}}(\phi) \le \overline{\Gamma}_j^{\mathrm{MMD}} < \infty. \quad (39)$$

**Proposition 9** (Finite-strata MMD learning bound)**.** Let $\delta_{\mathrm{D}} \in (0,1)$ and suppose

$$n_{\mathrm{D}} \ge \frac{8}{p_{\min}\eta} \log\left(\frac{8 M_j N_j^{\mathrm{fs}}}{\delta_{\mathrm{D}}}\right). \quad (40)$$

Define $e_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) \triangleq \rho_j \left[2 + \sqrt{2 \log\left(\frac{4 M_j N_j^{\mathrm{fs}}}{\delta_{\mathrm{D}}}\right)}\right] \sqrt{\frac{2}{n_{\mathrm{D}} p_{\min} \eta}}$; $r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) \triangleq 2 e_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) + 2\rho_j \sqrt{\frac{2\left[M_j \log 2 + \log\{4 N_j^{\mathrm{fs}} / \delta_{\mathrm{D}}\}\right]}{n_{\mathrm{D}}}}$. Then, with probability at least $1 - \delta_{\mathrm{D}}$,

$$\sup_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} \left|\widehat{D}_j^{\mathrm{MMD}}(\phi) - D_j^{\mathrm{MMD}}(\phi)\right| \le r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}). \quad (41)$$

If $\widehat{\phi}_j^{\mathrm{fs}}$ obeys $\widehat{D}_j^{\mathrm{MMD}}(\widehat{\phi}_j^{\mathrm{fs}}) \le \inf_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} \widehat{D}_j^{\mathrm{MMD}}(\phi) + \varepsilon_{\mathrm{opt},j}^{\mathrm{fs}}$, then, on the same event,

$$D_j^{\mathrm{MMD}}(\widehat{\phi}_j^{\mathrm{fs}}) \le \inf_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} D_j^{\mathrm{MMD}}(\phi) + 2 r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) + \varepsilon_{\mathrm{opt},j}^{\mathrm{fs}}, \quad (42)$$

$$|\tau_{\widehat{Z}_j^{\mathrm{fs}}} - \tau| \le \overline{\Gamma}_j^{\mathrm{MMD}} \left[\inf_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} D_j^{\mathrm{MMD}}(\phi) + 2 r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) + \varepsilon_{\mathrm{opt},j}^{\mathrm{fs}}\right], \quad (43)$$

where $\widehat{Z}_j^{\mathrm{fs}} = \widehat{\phi}_j^{\mathrm{fs}}(V_j)$, provided the primitive structural conditions and overlap used in the exact bias decomposition hold for every candidate.

*Proof.* Eq. (40) and a multiplicative Chernoff bound imply that every candidate treatment-by-stratum cell contains at least $n_{\mathrm{D}} p_{\min} \eta / 2$ observations except on an event of probability at most $\delta_{\mathrm{D}}/4$. Conditional on the cell assignments, bounded Hilbert-space mean concentration and a union bound over at most $2M_j N_j^{\mathrm{fs}}$ cells give $\left|\widehat{\Delta}_{j,z}^{\mathrm{MMD}}(\phi) - \Delta_{j,z}^{\mathrm{MMD}}(\phi)\right| \leq 2e_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}})$ simultaneously, with failure probability at most $\delta_{\mathrm{D}}/2$. A multinomial $L_1$ concentration bound gives $\sum_z |\widehat{p}_{j,z}(\phi) - p_{j,z}(\phi)| \leq \sqrt{\frac{2\left[M_j \log 2 + \log\{4N_j^{\mathrm{fs}}/\delta_{\mathrm{D}}\}\right]}{n_{\mathrm{D}}}}$ simultaneously over candidates, with failure probability at most $\delta_{\mathrm{D}}/4$. On the intersection of these events, $0 \leq \Delta_{j,z}^{\mathrm{MMD}}(\phi) \leq 2\rho_j$ and the triangle inequality give Eq. (41).

The empirical-risk comparison yields $D_j^{\mathrm{MMD}}(\widehat{\phi}_j^{\mathrm{fs}}) \leq \widehat{D}_j^{\mathrm{MMD}}(\widehat{\phi}_j^{\mathrm{fs}}) + r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) \leq \inf_{\phi \in \mathcal{F}_j^{\mathrm{fs}}} D_j^{\mathrm{MMD}}(\phi) + 2r_{j,n_{\mathrm{D}}}^{\mathrm{MMD}}(\delta_{\mathrm{D}}) + \varepsilon_{\mathrm{opt},j}^{\mathrm{fs}}$, which proves Eq. (42). The exact bias decomposition and the MMD stability condition in Eq. (39) give $|\tau_{Z_j^{\phi}} - \tau| \leq \Gamma_j^{\mathrm{MMD}}(\phi) D_j^{\mathrm{MMD}}(\phi)$. Apply this inequality to $\widehat{\phi}_j^{\mathrm{fs}}$ and substitute Eq. (42) to obtain Eq. (43). □