Joint Confounder Selection for Causal Mediation Analysis in High Dimensions

Chanmin Kim*
SungKyunKwan University
chanmin.kim@skku.edu

Hyunwoo Kim
University of Connecticut
jnp26001@uconn.edu

Abstract

The estimation of causal effects in observational research fundamentally relies on proper adjustment for confounding variables. While data-driven confounder selection is well-established for conventional exposure-outcome analyses, such methodologies remain underdeveloped in causal mediation analysis. This paper proposes a Bayesian nonparametric framework for confounder selection in mediation analysis using Bayesian Additive Regression Trees (BART). We introduce a shared sparsity-inducing prior across the exposure, mediator, and outcome models to identify a sufficient adjustment set that satisfies the mediation disjunctive cause criterion. Importantly, we provide rigorous theoretical guarantees for the proposed method: we establish the posterior consistency of confounder selection even in high-dimensional regimes ($P \gg N$) and demonstrate that this selection consistency translates into the consistent estimation of natural direct and indirect effects. The proposed method consistently outperforms competing approaches across a range of simulation scenarios, offering a principled and practical approach for high-dimensional mediation analysis.

1 Introduction

In observational studies, accurate estimation of causal effects requires careful adjustment for confounding variables that simultaneously influence both the exposure and the outcome. In modern applications, however, researchers often collect a large number of pre-treatment covariates without prior knowledge of which are true confounders. In such cases, identifying an appropriate subset of confounders is an important first step before applying causal inference methods. A growing body of research has addressed confounder selection. Loh and Vansteelandt [2021] proposed identifying an optimal set of confounders based on covariates included in propensity score models, with the aim of achieving stable causal effect estimates. Häggström [2018] used Markov and Bayesian networks for this purpose, while Shortreed and Ertefaie [2017] introduced the outcome-adaptive LASSO to select confounders for inclusion in propensity score models. Shi et al. [2019] employed neural networks to identify confounders relevant to both the propensity score and outcome models. In the Bayesian framework, Wang et al. [2012], Lefebvre et al. [2014], and Wang et al. [2015] developed inference methods for variable selection in propensity score and outcome models.

When a mediator lies on the causal pathway between an exposure and an outcome, mediation analysis decomposes the total effect into direct and indirect effects, offering insight into causal mechanisms. Unlike standard causal analyses that adjust for confounders between treatment and outcome, mediation analysis must address three sets of confounders: those affecting the treatment-mediator ($A - M$), treatment-outcome ($A - Y$), and mediator-outcome ($M - Y$) relationships, as shown in Figure 1 (a). To address these challenges, several assumptions and methodological

*Corresponding author: Chanmin Kim, Department of Statistics, SungKyunKwan University, South Korea

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

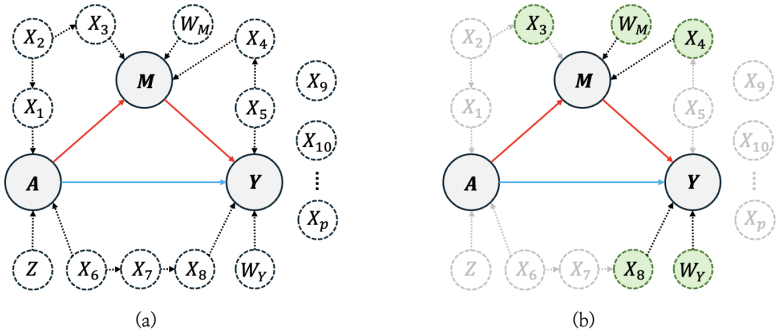

Figure 1: Directed acyclic graph (DAG) for mediation analysis. The red solid path (through the mediator M) denotes the indirect effect, and the blue solid path denotes the direct effect; dashed lines indicate confounding relationships among A, M, and Y. (a) A high-dimensional setting with candidate covariates $X_1, \ldots, X_p$ along with auxiliary variables Z, W_M, and W_Y. (b) The subset $\{X_3, W_M, X_4, X_8, W_Y\}$ (highlighted in green) satisfies the proposed *mediation disjunctive cause criterion* and is sufficient for identifying the natural direct and indirect effects.

frameworks have been proposed; detailed discussions can be found in Robins and Greenland [1992], Imai et al. [2010b], Pearl [2022].

In mediation analysis, selecting an appropriate set of confounders from a high-dimensional pool of candidate covariates is a crucial preliminary step. This task is particularly important and must be conducted with great care, as confounding bias can arise simultaneously across three distinct relationships. In general, confounders are selected based on domain knowledge or prior studies. However, to our knowledge, data-driven confounder selection methods specifically tailored for mediation analysis have not yet been thoroughly explored.

In this research, we develop new criteria for confounder selection specifically designed for mediation analysis. These criteria guide the identification of confounder sets that align with standard identification assumptions, such as sequential ignorability, especially in the presence of high-dimensional covariates. Building on these criteria, we propose a data-driven confounder selection method based on a Bayesian decision tree framework. Specifically, we place separate Bayesian additive regression trees (BART) priors on the treatment, mediator, and outcome models, with a common prior on the variable selection probabilities to identify variables consistently important across all three (i.e., the confounders). Together, these form an integrated framework with four contributions:

(*A new criterion*). We introduce the mediation disjunctive cause criterion, designed for the three-pathway confounding structure of mediation analysis.

(*Identification*). Theorem 3.2 establishes that any covariate set satisfying this criterion is a sufficient adjustment set under sequential ignorability.

(*A Bayesian nonparametric model*). We develop Bayesian causal mediation forests with a joint sparsity-inducing prior linking the exposure, mediator, and outcome models.

(*Consistency guarantees*). Theorems 6.1 and 6.3 and Corollary 6.2 establish posterior consistency for both confounder selection and causal-effect estimation, including the regime $P \gg N$.

Related work Vander Weele and Shpitser [2011] and VanderWeele [2019] proposed criteria for confounder selection in standard treatment-outcome analyses. In mediation analysis, however, confounding must be addressed across three distinct causal pathways (Figure 1), and clear, explicit criteria for confounder selection have yet to be established - a gap we seek to address. Our work also extends Kim et al. [2023], who used Bayesian additive regression trees (BART; Chipman et al., 2010) with a common prior on selection probabilities to identify confounders in treatment-outcome analyses. In contrast, we develop a method based on Bayesian causal mediation forests, a variant of Bayesian causal forests [Hahn et al., 2020], that jointly models the treatment, mediator, and outcome. Cheng et al. [2022] and Xu et al. [2023] also studied confounding in mediation analysis, but under the assumption that confounders are latent and recoverable only through observed proxies; our framework

instead targets the directly observed high-dimensional setting in which confounders themselves must be selected from the candidate covariates.

2 Notation and estimands

Beyond estimating causal effect sizes, causal mediation analysis provides insight into the underlying mechanisms by decomposing the total effect of an exposure (A) on an outcome (Y) into distinct components [Imai et al., 2010a, VanderWeele, 2015]. As shown in Figure 1, the total effect is partitioned into an *indirect effect* (red), which operates through a mediator (M), and a *direct effect* (blue), which captures the influence of A on Y not mediated by M. This decomposition enables attribution of the outcome variation to specific causal pathways, offering a more granular understanding of the relationship between A and Y. Let A_i, M_i, and Y_i denote the observed exposure, mediator, and outcome, respectively, for individual i, and let $\boldsymbol{X}_i$ be a P-dimensional vector of pre-treatment covariates. In the potential outcomes framework [Rubin, 1974], $M_i(a)$ denotes the potential mediator value that would be observed for unit i under exposure level a, and $Y_i(a, m)$ denotes the potential outcome under exposure a and mediator value m. Under the extended Stable Unit Treatment Value Assumption (SUTVA) for mediation [Forastiere et al., 2016], which assumes (1) no interference between units and (2) no multiple versions of treatment, the observed variables can be expressed in terms of potential outcomes as: $M_i = M_i(A_i)$ and $Y_i = Y_i(A_i, M_i(A_i))$.

A key challenge in mediation analysis is identifying unobserved potential outcomes such as $M_i(1 - A_i)$, $Y_i(1 - A_i, M_i(1 - A_i))$, and the cross-world quantity $Y_i(a, M_i(a'))$ for $a \neq a'$. Imai et al. [2010b] address this using the sequential ignorability assumption:

$$\{Y_i(a', m), M_i(a)\} \perp A_i \mid \boldsymbol{X}_i = \boldsymbol{x}, \tag{1}$$

$$Y_i(a', m) \perp M_i(a) \mid A_i = a, \boldsymbol{X}_i = \boldsymbol{x}, \tag{2}$$

for all $a, a' \in \{0, 1\}$, along with positivity $0 < \Pr(A_i = a \mid \boldsymbol{X}_i = \boldsymbol{x}), 0 < p(M_i(a) = m \mid A_i = a, \boldsymbol{X}_i = \boldsymbol{x})$ for all $m \in \mathcal{M}$ and $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$.

Assumption (1) requires adjustment for exposure-mediator and exposure-outcome confounders, while Assumption (2) pertains to mediator-outcome confounders unaffected by exposure. A unified criterion for selecting covariates satisfying both remains underexplored; we address this in the next section.

Under these assumptions, $Y_i(a, M_i(a'))$ is identifiable. For units with covariates $\boldsymbol{X}_i = \boldsymbol{x}$, we define:

$$\zeta(\boldsymbol{x}) = \mathbb{E}[\, Y_i(1, M_i(1)) - Y_i(1, M_i(0)) \mid \boldsymbol{X}_i = \boldsymbol{x}\,],$$
$$\delta(\boldsymbol{x}) = \mathbb{E}[\, Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \mid \boldsymbol{X}_i = \boldsymbol{x}\,].$$

Here, $\zeta(\boldsymbol{x})$ denotes the *indirect effect*, representing the effect mediated through M_i, while $\delta(\boldsymbol{x})$ denotes the *direct effect*, not operating through the mediator. Their sum yields the *total effect*, and averaging over $\boldsymbol{X}_i$ gives the marginal direct and indirect effects.

3 Mediation specific confounder selection criterion

Confounder selection is fundamental to causal inference. While domain-driven selection is common, it becomes impractical in high-dimensional settings, prompting the development of formal criteria for exposure-outcome analyses. Notable examples include: (a) the pretreatment criterion, which adjusts for all covariates measured prior to treatment; (b) the common cause criterion, which adjusts only for variables affecting both exposure and outcome; (c) the disjunctive cause criterion, which includes any variable that is a cause of the exposure, the outcome, or both; (d) in addition, the recommendation to exclude instrumental variables to avoid Z-bias amplification [VanderWeele, 2019].

However, these criteria are not designed for causal structures involving a mediator. To our knowledge, no formal criterion has been proposed for selecting confounders in mediation analysis. We address this gap by introducing a novel criterion that explicitly incorporates the mediator and supports identification under the sequential ignorability assumption.

Definition 3.1 (**Mediation Disjunctive Cause Criterion).** Let Π denote the set of all backdoor pathways between any pair of $\{A, M, Y\}$, i.e., paths that begin with an edge pointing into the first variable [Pearl, 2009] and are not already blocked by an unconditioned collider, and let $\mathrm{pa}(\cdot)$ denote

the set of direct causes. A set $\mathcal{C}$ satisfies the criterion if, for every backdoor pathway $\pi \in \Pi$, it holds that:

$$\mathcal{C} \cap \pi \cap \big(\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y)\big) \neq \emptyset,$$

subject to the constraint that $\mathcal{C}$ excludes any variable acting solely as an instrument for A.

In other words, a covariate set $\mathcal{C}$ constitutes a sufficient adjustment set if it ensures that every confounding pathway is intercepted by at least one direct cause of A, M, or Y, while strictly excluding pure instruments. Based on this definition, we establish the following result:

Theorem 3.2 (**Sufficiency of the Mediation Disjunctive Cause Criterion).** *Let $\boldsymbol{X}$ be the set of high-dimensional observed pre-treatment covariates. Suppose that every backdoor pathway $\pi \in \Pi$ contains at least one observed variable that is a direct cause of A, M, or Y, i.e., $\pi \cap \boldsymbol{X} \cap \big(\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y)\big) \neq \emptyset$. Then, there exists a subset $\mathcal{C} \subseteq \boldsymbol{X}$ that satisfies the mediation disjunctive cause criterion and constitutes a sufficient adjustment set for the sequential ignorability assumptions in Eqs. (1) and (2).*

The proof is provided in Appendix A. To illustrate the practicality of our assumption, consider a confounding pathway $A \leftarrow X_1 \leftarrow X_2 \rightarrow X_3 \rightarrow M$ in Figure 1(a). To block this backdoor path, one does not need to observe the entire set $\{X_1, X_2, X_3\}$; a set satisfying the mediation disjunctive cause criterion only needs one direct cause along each backdoor path, supplemented by '*additional predictors*' of M and Y (Figure 1 (b)). More generally, whenever the observed covariates are rich enough to d-separate all backdoor paths among the exposure, mediator, and outcome [Pearl, 2009], valid adjustment sets $\mathcal{C}$ satisfying sequential ignorability can be constructed. In the next section, we propose a data-driven method to identify one such set $\mathcal{C}$ by favoring a sparse selection of covariates.

4 BART for causal mediation analysis

To identify confounders in high-dimensional mediation settings, we adopt a Bayesian nonparametric approach based on Bayesian Additive Regression Trees (BART) [Chipman et al., 2010]. BART models the outcome as a sum over T regression trees:

$$Y_i = f(\boldsymbol{X}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad f(\boldsymbol{X}_i) = \sum_{t=1}^{T} g(\boldsymbol{X}_i; \mathcal{T}_t, \mathcal{M}_t),$$

where each $g(\cdot)$ is a regression tree with structure $\mathcal{T}_t$ and terminal node parameters $\mathcal{M}_t$. Each tree comprises internal nodes consisting of splitting rules of the form $X_j < c$, where X_j is a selected splitting variable from the pool. Trees are typically shallow and act as weak learners. BART updates tree structures via Bayesian backfitting [Hastie and Tibshirani, 2000], iteratively modifying each tree based on residuals from the others. Tree updates randomly apply one of four alterations: grow, prune, change, or swap. In grow and change steps, splitting variables are drawn according to a probability vector $\boldsymbol{s} = (s_1, \ldots, s_P)$ over the P covariates. While a uniform prior is common for $\boldsymbol{s}$, sparsity-inducing priors improve variable selection [Linero, 2018]. We adopt a symmetric Dirichlet prior: $\boldsymbol{s} \sim \mathrm{Dir}\left(\frac{\alpha}{P}, \ldots, \frac{\alpha}{P}\right)$, which yields a simple conjugate update and encourages concentrated variable selection, promoting the identification of key confounders.

4.1 Observed data model for mediation

Extending the BART framework, we model the exposure, mediator, and outcome as

$$\Pr(A_i = 1) = \Phi\{\mu(\boldsymbol{X}_i)\}, \tag{3}$$

$$M_i = \mu_M(\boldsymbol{X}_i, \hat{\pi}_i) + \tau_M(\boldsymbol{X}_i) A_i + \epsilon_i, \tag{4}$$

$$Y_i = \mu_Y(\boldsymbol{X}_i, \hat{\pi}_i, \hat{m}_{0i}, \hat{m}_{1i}) + \tau_Y(\boldsymbol{X}_i) A_i + \tau'_Y(\boldsymbol{X}_i) M_i + \eta_i, \tag{5}$$

where $\Phi(\cdot)$ is the standard normal CDF, $\epsilon_i \sim \mathcal{N}(0, \sigma_M^2)$, and $\eta_i \sim \mathcal{N}(0, \sigma_Y^2)$. Independent BART priors are placed on $\mu, \mu_M, \tau_M, \mu_Y, \tau_Y, \tau'_Y$. The mediator model follows the Bayesian Causal Forest (BCF) of Hahn et al. [2020], where $\tau_M(\boldsymbol{x})$ captures the conditional effect of exposure on the mediator and $\mu_M(\boldsymbol{x}, \cdot)$ adjusts for confounding; the outcome model builds on the Bayesian Causal Mediation Forest (BCMF) of Ting and Linero [2025], with $\tau_Y(\boldsymbol{x})$ representing the conditional direct effect $\delta(\boldsymbol{x})$ and the product $\tau'_Y(\boldsymbol{x}) \cdot \tau_M(\boldsymbol{x})$ corresponding to the conditional indirect effect $\zeta(\boldsymbol{x})$.

To mitigate *regularization-induced confounding* [Hahn et al., 2020], also referred to as *prior dogmatism* [Linero, 2024], the prognostic functions μ_M and μ_Y are augmented with the estimated propensity score $\hat{\pi}_i = \hat{\mathbb{E}}(A_i \mid \boldsymbol{X}_i)$ and, following Ting and Linero [2025], with the clever covariates $\hat{m}_{ai} = \hat{\mathbb{E}}(M_i \mid A_i = a, \boldsymbol{X}_i)$ for $a \in \{0, 1\}$. For prognostic functions (μ, μ_M, μ_Y) we use a relatively large number of trees (e.g., 200) to capture complex covariate relationships, while modifier functions $(\tau_M, \tau_Y, \tau'_Y)$ use fewer trees (e.g., 20) with priors that shrink toward homogeneity to discourage unnecessary variable inclusion.

4.2 Joint confounder selection via reparameterized BART priors

To perform confounder selection across the exposure, mediator, and outcome models in Eqs. (3), (4), and (5), we propose a unified approach linking the prognostic functions μ, μ_M, and μ_Y. Each function is modeled with a BART prior, where internal nodes split on covariates drawn according to a selection probability vector $\boldsymbol{s}$, a parameter of the prior rather than of Eqs. (3)–(5). By assigning a common sparsity-inducing Dirichlet prior to the selection probabilities, we jointly identify covariates that are consistently important for confounding adjustment (i.e., true confounders).

In Models (4) and (5), additional variables $(\hat{\pi}_i, \hat{m}_{0i}, \hat{m}_{1i})$ are included as inputs, each with associated selection probabilities (s_π, s_{m0}, s_{m1}). As such, the three models have nested input sets of sizes P, $P + 1$, and $P + 3$, and hence distinct but overlapping selection vectors: $\boldsymbol{s}^A, \boldsymbol{s}^M, \boldsymbol{s}^Y$, making a naïve prior specification nontrivial. We resolve this by reparameterizing all three vectors in terms of a common selection vector $\boldsymbol{s}^Y = \boldsymbol{s} = (s_1, \ldots, s_P, s_\pi, s_{m0}, s_{m1})$, and defining:

$$
\begin{aligned}
\boldsymbol{s}^A &= \left(\frac{s_1}{1 - (s_\pi + s_{m0} + s_{m1})}, \ldots, \frac{s_P}{1 - (s_\pi + s_{m0} + s_{m1})} \right), \\
\boldsymbol{s}^M &= \left(\frac{s_1}{1 - (s_{m0} + s_{m1})}, \ldots, \frac{s_P}{1 - (s_{m0} + s_{m1})}, \frac{s_\pi}{1 - (s_{m0} + s_{m1})} \right), \\
\boldsymbol{s}^Y &= (s_1, \ldots, s_P, s_\pi, s_{m0}, s_{m1}).
\end{aligned}
$$

This reparameterization leverages the neutrality property of the Dirichlet distribution: dropping a subset of coordinates and renormalizing the rest yields again a Dirichlet, so $\boldsymbol{s}^A$ and $\boldsymbol{s}^M$ are the conditionals of $\boldsymbol{s}$ on the coordinates each model uses. We impose the prior $\boldsymbol{s} \sim \text{Dir}(\alpha/(P + 3), \ldots, \alpha/(P + 3))$, which induces consistent sparsity across all models.

Given this formulation, the posterior update of $\boldsymbol{s}$ is proportional to:

$$
\begin{aligned}
\mathcal{Q} = \prod_{t=1}^{T} & \left\{ \prod_{b \in \mathcal{T}_t^\mu} \frac{s_{j_b}}{1 - (s_\pi + s_{m0} + s_{m1})} \times \prod_{c \in \mathcal{T}_t^{\mu_M}} \frac{s_{j_c}}{1 - (s_{m0} + s_{m1})} \times \prod_{e \in \mathcal{T}_t^{\mu_Y}} s_{j_e} \right\} \\
& \times \left\{ (s_\pi s_{m0} s_{m1})^{\alpha/(P+3)-1} \cdot \prod_{j=1}^{P} s_j^{\alpha/(P+3)-1} \right\},
\end{aligned} \tag{6}
$$

where $\mathcal{T}_t^\mu$, $\mathcal{T}_t^{\mu_M}$, and $\mathcal{T}_t^{\mu_Y}$ denote the sets of internal nodes in trees associated with μ, μ_M, and μ_Y, respectively. The sparsity-inducing prior favors a sparse subset among the covariate sets $\mathcal{C}$ that satisfy the mediation disjunctive cause criterion. In doing so, it identifies variables with the strongest signals along backdoor paths and thus selects direct causes rather than upstream ancestors (e.g., X_3 rather than X_2 on the A-M path in Figure 1(b)). An empirical demonstration of this behavior is provided in Appendix F (Table 3 and Figure 6).

5 Posterior computation

The posterior is computed via a Gibbs sampler consisting of two main blocks. In the first block, Bayesian backfitting [Hastie and Tibshirani, 2000] is applied to each of the six BART components $(\mu, \mu_M, \tau_M, \mu_Y, \tau_Y, \tau'_Y)$: for each tree in turn, a partial residual is formed by removing the tree's contribution, and its structure and leaf parameters are updated via Metropolis–Hastings proposals (grow/prune/change). The probit exposure model is handled by data augmentation [Albert and Chib, 1993]: latent $Z_i = \mu(\boldsymbol{X}_i) + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$, are drawn from a normal truncated to the sign of A_i, after which μ is updated as in a Gaussian regression. The residual variances σ_M^2 and σ_Y^2 are then

drawn from their conjugate inverse-gamma full conditionals. In the second block, the shared selection probability vector $\boldsymbol{s}$ is updated via a Metropolis–Hastings step targeting the joint posterior in Eq. (6); this is the key step that couples variable selection across the exposure, mediator, and outcome models. The full sampling procedure is outlined in Algorithm 1 of Appendix D.

6 Theoretical properties

Let $\mathcal{D} = \{(A_i, M_i, Y_i, \boldsymbol{X}_i)\}_{i=1}^N$ denote the observed dataset, where A_i is the exposure, M_i the mediator, Y_i the outcome, and $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iP})$ is a P-dimensional vector of covariates. Assuming that all direct causes of M other than A, and of Y other than A and M, are observed, we define $\mathcal{C}^\star = \mathrm{pa}(M) \cup \mathrm{pa}(Y)$, the set of all direct causes of the mediator and the outcome and nothing else (minimal in this sense, not in cardinality). Under this assumption $\mathcal{C}^\star$ satisfies the mediation disjunctive cause criterion and is therefore a sufficient adjustment set (Corollary A.1).

Theorem 6.1 (**Posterior Consistency of Confounder Selection).** *Let $\mathcal{C}^\star$ be the covariate set defined above, and let $\mathcal{S}$ be the model index, i.e., the subset of covariates on which the trees may split. Under the regularity and support conditions of Appendix B (Assumptions A–C), if the concentration parameter of the sparsity-inducing Dirichlet prior satisfies $\alpha_N \asymp N^{-\gamma}$ for a sufficiently large $\gamma > 0$, then as sample size $N \to \infty$, the posterior probability of selecting the true set $\mathcal{C}^\star$ converges to 1:*

$$P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1.$$

Proof. The proof is provided in Appendix B. □

Corollary 6.2 (**High-Dimensional Consistency).** *The consistency result in Theorem 6.1 extends to the high-dimensional setting where the number of covariates P_N grows with the sample size N (potentially $P_N \gg N$), provided that the growth rate satisfies $\log P_N = o(N)$. If the Dirichlet prior satisfies Assumption B with P replaced by P_N, so that each additional active covariate costs a prior factor of order $N^{-\gamma} P_N^{-1}$, then the posterior consistency holds:*

$$P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1,$$

for a sufficiently large $\gamma > 0$.

Proof. The proof is provided in Appendix B. □

Theorem 6.3 (**Posterior Consistency of Natural Direct and Indirect Effects).** *Let $\Psi_0 = (\delta_0, \zeta_0)$ denote the true natural direct and indirect effects, and let $\Psi = (\delta, \zeta)$ denote the random variables corresponding to these effects derived from the posterior distribution. Under Assumptions A–C (Appendix B) and the result of Theorem 6.1, the posterior distribution of the estimated effects concentrates around the true values. That is, for any $\xi > 0$,*

$$P(\|\Psi - \Psi_0\|_2 > \xi \mid \mathcal{D}) \xrightarrow{P} 0 \quad \text{as } N \to \infty,$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm for vectors.

Proof. The proof is provided in Appendix B. □

Remark 6.4. Together, Theorems 6.1 and 6.3 and Corollary 6.2 provide a complete theoretical guarantee for the proposed framework: Theorem 6.1 establishes consistent selection of the target set $\mathcal{C}^\star$, Corollary 6.2 extends this to the high-dimensional regime ($P \gg N$) by showing that the sparsity-inducing prior counteracts the combinatorial explosion of candidate models, and Theorem 6.3 translates the resulting posterior contraction of the regression functions into consistency of the natural direct and indirect effect estimates, yielding the oracle property asymptotically.

7 Experiments

We evaluate the proposed model across simulation scenarios with $P = 100$ covariates and sample size $N = 250$, with $X_1 - X_5$ designated as true confounders. The settings cover heterogeneous treatment effects, non-linear mediator/outcome relationships, instrumental variables, and non-confounding predictors: **Scenario 1**: Heterogeneous effects with non-linear mediator and outcome models; **Scenario 2**: Adds two non-confounding predictors to the outcome model; **Scenario 3**: Adds two instruments to the exposure model; **Scenario 4**: Adds two instruments to the mediator model on top of Scenario 3; **Scenarios 5-6**: High-dimensional ($P = 500$ and $P = 2000$); otherwise as Scenario 3; **Scenario 7**: Low-dimensional ($P = 20$); otherwise as Scenario 3. Full data-generating mechanisms are in Appendix E.

We compare our method (**BCMF-CS**; Bayesian causal mediation forest with confounder selection) against three baselines: (1) **BCMF**, our model with uniform (non-sparse) selection priors [Ting and Linero, 2025]; (2) **DCMF** (Dirichlet causal mediation forest, named after DART; Linero, 2018), with independent sparsity-inducing Dirichlet priors per model component (no joint selection); and (3) **LSEM**, a classical linear structural equation model [Baron and Kenny, 1986, MacKinnon and Dwyer, 1993, Sobel, 1996] fit with the five true confounders plus five additional covariates (a slight advantage, since LSEM lacks variable selection). All methods are evaluated over $S = 200$ replicates; full specifications are in Appendix C. Performance is assessed along two dimensions. For *effect estimation*, we report per-unit bias, absolute bias, MSE, and coverage, averaged over replicates; e.g., Bias $= \frac{1}{S}\sum_{s=1}^{S}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\zeta(\boldsymbol{X}_i) - \hat{\zeta}^{(s)}(\boldsymbol{X}_i)\right)\right\}$. For *confounder selection*, we report TPR, FPR, PPV, and F1, treating a variable as selected if its posterior inclusion probability (PIP) exceeds 0.5 [Barbieri and Berger, 2004]; comparisons against the ground truth are made separately for the exposure, mediator, and outcome models.

Table 1 reports the results for effect estimation. Across all scenarios, BCMF-CS consistently achieves superior performance in estimating causal effects. While DCMF shows similar performance in some scenarios, its accuracy notably deteriorates in Scenarios 2 and 3 when non-confounding predictors or instruments are introduced. In these settings, BCMF-CS outperforms DCMF in terms of absolute bias, MSE, and coverage. These improvements stem from BCMF-CS's ability to accurately identify true confounders.

Table 7 (Appendix I) summarizes confounder selection performance. Competing models such as BCMF and DCMF do not effectively select variables for the exposure model (A), resulting in lower performance in this component. Even in the mediator (M) and outcome (Y) models (more directly tied to effect estimation), BCMF-CS consistently outperforms all baselines across evaluation metrics.

Figures 4 and 5 (Appendix I) display posterior inclusion probabilities under Scenarios 2 and 4, which respectively introduce additional outcome predictors and mediator instruments. BCMF-CS reliably selects the true confounders with high probability while suppressing irrelevant variables. In Figure 4, the additional outcome predictors X_6 and X_7 are selected with high PIP, while in Figure 5, the instruments X_8 and X_9 are excluded from the outcome model-illustrating that BCMF-CS respects the mediation disjunctive cause criterion.

Sensitivity analysis and additional experiments We assessed prior sensitivity (Scenario 1) by varying the number of trees for the prognostic functions (μ, μ_M, μ_Y) from the default 200 to 100 and 50, while keeping 20 trees for the modifier functions (τ_Y, τ'_Y). Table 2 shows stable performance once a sufficient number of trees is used, consistent with Chipman et al. [2010].

We further conducted **additional experiments**:

(1) Sensitivity to the Dirichlet concentration parameter,

(2) The impact of unmeasured confounding (two settings),

(3) The effect of model misspecification,

(4) Identification of direct causes under multiple ancestors in the backdoor path,

(5) Dependent covariates replicating the DAG of Figure 1(b), including a stress variant with an ancestor as strongly associated with M as the direct parent,

(6) Comparison with an external confounder-selection baseline (outcome-adaptive LASSO), and

Table 1: Results (bias, absolute bias, mean squared error, and coverage) for different scenarios: Scenario 1 (base), Scenario 2 (with two predictors), Scenarios 3 and 4 (with two instruments in exposure and mediator models, respectively), Scenario 5 ($N < P = 500$), Scenario 6 ($N < P = 2000$), and Scenario 7 ($P = 20$). Results are reported separately for the direct effect (DE) and indirect effect (IE).

		BIAS		Abs.BIAS		MSE		Coverage	
Scenario	Method	DE	IE	DE	IE	DE	IE	DE	IE
(1)	BCMF-CS	0.048	-0.024	0.172	0.248	0.056	0.167	0.99	0.90
	BCMF	0.128	0.016	0.308	0.447	0.179	0.415	0.92	0.69
	DCMF	0.023	-0.001	0.170	0.338	0.052	0.259	0.98	0.76
	LSEM	1.091	0.213	1.100	1.075	1.438	1.717	0.19	0.39
(2)	BCMF-CS	0.009	0.025	0.379	0.409	0.251	0.346	0.87	0.72
	BCMF	0.099	0.133	0.769	0.856	0.906	1.159	0.64	0.18
	DCMF	-0.044	0.112	0.774	0.788	0.937	1.055	0.61	0.26
	LSEM	1.001	0.210	1.110	1.068	1.878	1.695	0.55	0.39
(3)	BCMF-CS	-0.038	0.038	0.308	0.488	0.167	0.454	0.77	0.59
	BCMF	-0.028	0.123	0.599	0.839	0.464	1.123	0.52	0.17
	DCMF	-0.098	0.119	0.704	0.811	0.785	1.096	0.47	0.22
	LSEM	0.928	0.158	0.947	1.047	1.121	1.649	0.54	0.39
(4)	BCMF-CS	-0.014	0.036	0.144	0.390	0.050	0.323	0.94	0.72
	BCMF	0.019	0.110	0.464	0.763	0.307	0.969	0.70	0.31
	DCMF	-0.011	0.031	0.211	0.434	0.133	0.411	0.87	0.67
	LSEM	0.945	0.196	0.960	1.069	1.149	1.704	0.53	0.39
(5)	BCMF-CS	-0.016	-0.021	0.087	0.332	0.022	0.234	0.99	0.80
	BCMF	0.217	0.035	0.402	0.993	0.463	1.596	0.85	0.18
	DCMF	-0.012	0.001	0.437	0.617	0.553	0.808	0.73	0.49
	LSEM	1.024	0.265	1.041	1.169	1.357	2.056	0.54	0.44
(6)	BCMF-CS	- 0.058	0.031	0.058	0.429	0.004	0.345	1.00	0.62
	BCMF	0.186	0.144	0.358	1.105	0.422	1.787	0.82	0.06
	DCMF	-0.122	0.132	0.878	1.004	0.956	1.486	0.36	0.02
	LSEM	0.944	0.067	1.104	1.062	1.820	1.765	0.52	0.34
(7)	BCMF-CS	0.001	-0.008	0.267	0.444	0.149	0.379	0.87	0.69
	BCMF	0.215	0.038	0.402	0.988	0.451	1.583	0.86	0.19
	DCMF	-0.017	-0.007	0.397	0.583	0.485	0.749	0.76	0.53
	LSEM	1.011	0.272	1.030	1.169	1.342	2.051	0.55	0.43

Table 2: Results (bias, absolute bias, mean squared error, and coverage) for sensitivity analysis with different numbers of trees.

Numbers of trees in	BIAS		Abs.BIAS		MSE		Coverage	
prognostic functions	DE	IE	DE	IE	DE	IE	DE	IE
50	0.046	-0.023	0.158	0.228	0.049	0.158	0.99	0.92
100	0.047	-0.032	0.149	0.228	0.049	0.169	0.98	0.92

(7) Interval widths underlying the coverage results.

The detailed results are provided in Appendix F.

Computational Efficiency. Figure 3 summarizes the runtime of BCMF-CS based on 5,000 MCMC iterations executed on a Mac Studio (Apple M1, 128 GB RAM). We varied the number of observations N and covariates P, using the tree configuration described in Appendix C. The runtime scales moderately with N, while remaining largely unaffected by P, reflecting the sparsity-inducing prior's effectiveness in focusing computation on relevant covariates. Overall, the method achieves practical scalability for moderate to large datasets.

8 Real Application: AIDS Clinical Trials Group Study 175

The ACTG175 study [Hammer et al., 1996] is a randomized clinical trial evaluating antiretroviral therapies for HIV-positive patients, specifically comparing zidovudine alone, didanosine alone, and their combination. The dataset, available in the R package `speff2trial`, includes 2,139 patients and 27 variables. The outcome is missing for 37.3% of patients, so we analyze the $N = 1{,}342$ complete cases, assuming missingness at random given the covariates. We define the outcome (Y) as CD4 T cell count at 95 ± 5 weeks and the mediator (M) as CD4 count at 20 ± 5 weeks. The binary exposure (A) distinguishes zidovudine only ($A = 0$) from the other regimens ($A = 1$). In addition, 15 pretreatment covariates (e.g., age, weight) are included, as detailed in Appendix G.

Figure 2 (left) displays estimated individual total, direct, and indirect effects. Direct effects exhibit limited heterogeneity, whereas indirect effects are clearly stratified by race (black circles: white; red triangles: non-white), suggesting that race contributes to heterogeneity in mediated pathways but not in direct effects. Race thus acts as an effect modifier, entering through τ_M and τ'_Y, rather than as a confounder entering through μ_M and μ_Y, which is why it is not selected in Figure 2 (right). Figure 2 (right) shows posterior inclusion probabilities (PIPs) for confounders across the exposure, mediator, and outcome models. Two variables-baseline CD4 count (`cd40`) and prior antiretroviral duration (`preanti`)-are consistently selected (PIP = 1). Because treatment is randomized but the mediator is not, only mediator–outcome confounding requires adjustment [VanderWeele, 2015]; cd40 and preanti are clinically plausible common causes of the two CD4 measurements, which we read as a plausibility check since the true confounders are unknown. Their PIPs of one in the exposure model do not mean that they predict treatment: all three models draw splitting variables from the common vector $\boldsymbol{s}$, so a variable selected for the mediator and outcome models is also split on in the exposure model, where such splits are harmless under randomization (Appendix I).

It is important to note that although the mediator serves as a short-term proxy for the outcome, long-term effects (e.g., over a 75-week period) can still emerge, particularly depending on exposure status. Because the indirect effect reflects both the $A \to M$ and $M \to Y$ pathways, it may account for only part of the total effect when the $A \to M$ association is weak, even if the $M \to Y$ association is strong. Consequently, a direct effect may also be present. Similar delayed responses have been documented in prior studies (see *Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents With HIV*, `https://clinicalinfo.hiv.gov/`).

A leave-one-out analysis (removing `cd40` or `preanti` in turn) yields nearly identical effect estimates, with `age` emerging as a newly selected confounder in either case. This is consistent with `age` being a parent of both variables: since our method targets direct causes of treatment, mediator, and outcome, it substitutes one sufficient blocking variable for another when a direct cause is removed. This behavior is precisely what the mediation disjunctive cause criterion guarantees-validity does not require a unique adjustment set, but any set that blocks all backdoor paths. Full results are reported in Appendix I (Figures 7 and 8).

9 Limitations and future directions

A fundamental assumption of our framework is that the observed covariate set is sufficient to satisfy sequential ignorability. While unmeasured confounding remains an inherent limitation of all observational studies, the mediation disjunctive cause criterion offers a practical safeguard: validity does not require the exhaustive enumeration of every direct cause, but rather at least one observed variable (or close proxy) along each backdoor path, which is an assumption that is plausible in high-dimensional settings with rich covariate information.

While our theoretical development focuses on the single-mediator setting, the proposed BART-based selection mechanism extends naturally to multiple or sequential mediators. We demonstrate this in Appendix H, where we outline the model modifications for a two-mediator scenario and provide validation results from a toy example. A fully general theoretical treatment is left for future work; these preliminary results confirm that our sparsity-inducing prior approach accommodates more complex mediation structures without fundamental changes to the algorithm.

The contribution of this paper goes beyond a new application of BART to variable selection. It consists of four interconnected advances: (1) the *mediation disjunctive cause criterion*, designed for the three-pathway confounding structure of mediation analysis; (2) Theorem 3.2, showing that any covariate

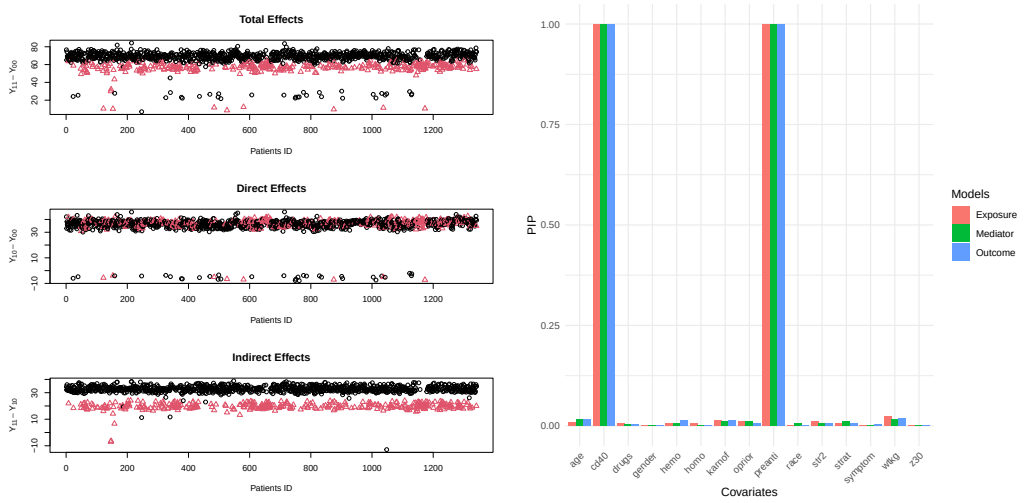

Figure 2: Results based on the ACTG175 dataset. (Left) Estimated individual total, direct, and indirect effects, ordered by patient ID. Red triangles represent non-white patients; black circles represent white patients. (Right) Posterior inclusion probabilities of potential confounders in the exposure, mediator, and outcome models.

set satisfying this criterion is a sufficient adjustment set under sequential ignorability; (3) a Bayesian causal mediation forest with a reparameterized joint sparsity-inducing prior that enforces this criterion in a data-driven way across the exposure, mediator, and outcome models; and (4) Theorems 6.1 and 6.3 together with Corollary 6.2, providing posterior consistency guarantees for both confounder selection and causal effect estimation, including the high-dimensional regime $P_N \gg N$. Together, these results yield a complete pipeline-from criterion formulation, through identification, to consistent estimation-that, to our knowledge, does not exist elsewhere in the mediation analysis literature.

Acknowledgements

The authors gratefully acknowledge that this work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (RS-2025-00554477, RS-2024-00407300).

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897, 2004.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

Lu Cheng, Ruocheng Guo, and Huan Liu. Causal mediation analysis with hidden confounders. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 113–122, 2022.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Laura Forastiere, Fabrizia Mealli, and Tyler J VanderWeele. Identification and estimation of causal mechanisms in clustered encouragement designs: disentangling bed nets using Bayesian principal stratification. *Journal of the American Statistical Association*, 111(514):510–525, 2016.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

Jenny Häggström. Data-driven confounder selection via markov and Bayesian networks. *Biometrics*, 74(2):389–398, 2018.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15): 1081–1090, 1996.

Trevor Hastie and Robert Tibshirani. Bayesian backfitting (with comments and a rejoinder by the authors. *Statistical Science*, 15(3):196–223, 2000.

Kosuke Imai and Teppei Yamamoto. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171, 2013.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010a.

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010b.

Adam Kapelner and Justin Bleich. bartmachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70:1–40, 2016.

Chanmin Kim, Mauricio Tec, and Corwin Zigler. Bayesian nonparametric adjustment of confounding. *Biometrics*, 79(4):3252–3265, 2023.

Geneviève Lefebvre, Joseph A Delaney, and Robyn L McClelland. Extending the Bayesian adjustment for confounding algorithm to binary treatment covariates to estimate the effect of smoking on carotid intima-media thickness: the multi-ethnic study of atherosclerosis. *Statistics in Medicine*, 33(16):2797–2813, 2014.

Antonio R Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.

Antonio R Linero. In nonparametric and high-dimensional models, Bayesian ignorability is an informative prior. *Journal of the American Statistical Association*, 119(548):2785–2798, 2024.

Wen Wei Loh and Stijn Vansteelandt. Confounder selection strategies targeting stable treatment effect estimators. *Statistics in Medicine*, 40(3):607–630, 2021.

David P MacKinnon and James H Dwyer. Estimating mediated effects in prevention studies. *Evaluation Review*, 17(2):144–158, 1993.

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

Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

Veronika Ročková and Stephanie Van der Pas. Posterior concentration for bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131, 2020.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Michael E Sobel. An introduction to causal inference. *Sociological Methods & Research*, 24(3): 353–379, 1996.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Angela Ting and Antonio R Linero. Estimating heterogeneous causal mediation effects with bayesian decision tree ensembles. *Journal of the American Statistical Association*, 120(551):1400–1413, 2025.

Tyler J Vander Weele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4): 1406–1413, 2011.

Tyler J VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

Tyler J VanderWeele. Principles of confounder selection. *European Journal of Epidemiology*, 34: 211–219, 2019.

Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–671, 2012.

Chi Wang, Francesca Dominici, Giovanni Parmigiani, and Corwin Matthew Zigler. Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3):654–665, 2015.

Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Ke Wang. Disentangled representation for causal mediation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10666–10674, 2023.

A Technical appendices and supplementary material

Appendix A: Proof of Theorem 3.2

Theorem 3.2 (**Sufficiency of the Mediation Disjunctive Cause Criterion).** *Let $\boldsymbol{X}$ be the set of high-dimensional observed pre-treatment covariates. Suppose that every backdoor pathway $\pi \in \Pi$ contains at least one observed variable that is a direct cause of A, M, or Y, i.e., $\pi \cap \boldsymbol{X} \cap \big(\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y)\big) \neq \emptyset$. Then, there exists a subset $\mathcal{C} \subseteq \boldsymbol{X}$ that satisfies the mediation disjunctive cause criterion and constitutes a sufficient adjustment set for the sequential ignorability assumptions in Eqs. (1) and (2).*

Proof. Let $\mathcal{C} \subseteq \boldsymbol{X}$ be the subset of observed covariates consisting of all variables in $\boldsymbol{X}$ that are direct causes (parents) of A, M, or Y. Formally, we define $\mathcal{C} = \boldsymbol{X} \cap (\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y))$. We first establish that $\mathcal{C}$ satisfies the sequential ignorability assumption for the assignment of A (Eq. 1). Consider any backdoor path from A to M or from A to Y. By the theorem's assumption, the high-dimensional set $\boldsymbol{X}$ contains at least one variable belonging to $\mathrm{pa}(A) \cup \mathrm{pa}(M) \cup \mathrm{pa}(Y)$ along every confounding pathway. This implies that the set of all observed covariates $\boldsymbol{X}$ is sufficient to block all backdoor paths originating from A. Based on Theorem 1 in Vander Weele and Shpitser [2011], if a sufficient adjustment set exists within the observed variables, then the subset consisting of the observed causes of the exposure, mediator, and outcome is also a sufficient adjustment set. (Their theorem is stated for a scalar outcome; its argument applies unchanged to the composite outcome $(M(a), Y(a', m))$ used here.) Therefore, $\mathcal{C}$, which includes all observed parents of A (and M, Y), blocks all backdoor paths from A. Thus, $\{Y(a', m), M(a)\} \perp A \mid \mathcal{C}$ holds.

Next, we address the ignorability condition for the mediator (Eq. 2), specifically $Y(a', m) \perp M(a) \mid A = a, \mathcal{C}$. Conditioning on $A = a$ blocks all paths passing through A. We must ensure that any

remaining backdoor paths from M to Y (not passing through A) are blocked. By the theorem's assumption, $\boldsymbol{X}$ intercepts every confounding pathway between M and Y at a parent node. Again, applying the disjunctive cause criterion in Vander Weele and Shpitser [2011], since $\mathcal{C}$ contains all observed direct causes of M and Y in $\boldsymbol{X}$, it is sufficient to block all backdoor paths between M and Y. (Here the theorem is applied to the pair (M, Y) within each stratum $A = a$, where conditioning on A closes all paths through A and only pre-treatment covariates remain on the backdoor paths.) Since $\mathcal{C}$ consists of pre-treatment covariates (non-descendants of A and M), conditioning on $\mathcal{C}$ does not open any collider paths. Consequently, $\mathcal{C}$ satisfies the mediation disjunctive cause criterion and ensures that sequential ignorability holds. □

Corollary A.1 (Sufficiency of $\mathcal{C}^\star$). If all direct causes of M other than A, and of Y other than A and M, are observed, then $\mathcal{C}^\star = \mathrm{pa}(M) \cup \mathrm{pa}(Y)$ satisfies the mediation disjunctive cause criterion and is a sufficient adjustment set.

Proof. Any backdoor path between a pair of $\{A, M, Y\}$ must reach its endpoint M or Y through a final edge $W \to M$ or $W \to Y$: a path exiting its endpoint through a descendant contains an unconditioned collider and is already blocked. Such a W is a direct parent of M or Y, observed by assumption, and hence in $\mathcal{C}^\star$, so every open backdoor path is intercepted at a direct cause; a pure instrument is a parent of neither M nor Y and is excluded automatically. Thus $\mathcal{C}^\star$ satisfies Definition 3.1, and Theorem 3.2 gives sufficiency. □

Note that $\mathcal{C}^\star$ is one particular criterion-satisfying set, distinct from the illustrative set in Figure 1(b); when some parents of M or Y are unobserved, Theorem 3.2 still yields a valid adjustment set within $\boldsymbol{X}$, possibly including parents of A.

Appendix B: Proofs of Theorems 6.1 and 6.3, and Corollary 6.2

In this section, we provide the regularity conditions and detailed proofs for the consistency of confounder selection. We consider the regime where $N > P$ (or P is fixed), allowing for the use of standard nonparametric sieve theory extended to generalized interaction models. Throughout the proofs, we distinguish between vector and function norms. For a vector $v \in \mathbb{R}^d$, $\|v\|_2$ denotes the standard Euclidean norm. For a function $f : \mathcal{X} \to \mathbb{R}$, $\|f\|_{L_2(P)}$ denotes the L_2 norm with respect to the marginal distribution of covariates P, defined as $\|f\|_{L_2(P)} = \left(\int_{\mathcal{X}} f(x)^2 dP(x)\right)^{1/2}$. This norm is intrinsic to the analysis of Kullback-Leibler divergence in regression settings.

Assumption A (Regularity and sieve construction).

1. Function space: The true prognostic and effect functions, denoted by f_0, are assumed to conform to a generalized additive structure with a maximum interaction order of $q \geq 1$. Specifically, f_0 consists of components where each depends on at most q variables. Each component function belongs to a Hölder ball $\mathcal{H}^\beta([0,1]^P, L)$ defined on the compact domain $\mathcal{X}$, with smoothness $0 < \beta \leq 1$ [Ghosal et al., 2000, Ročková and Van der Pas, 2020].
2. Design distribution: The covariates X_i are independent and identically distributed (i.i.d.) with a probability density $p(x)$ that is strictly bounded away from zero and infinity on $\mathcal{X}$ [Stone, 1982].
3. Sieve construction and tree prior: The parameter space is restricted to a sieve $\mathcal{F}_N$ of sum-of-trees models. To ensure the metric entropy is controlled and the posterior complexity adapts to the effective dimension, we impose the following constraints on the tree structures:
   (1) Each ensemble consists of binary decision trees.
   (2) The prior distribution on the total number of leaf nodes K satisfies an exponential decay condition:
   $$\pi(K) \leq C\lambda^{-K} = C\exp(-C'K),$$
   for constants $C > 0$ and $\lambda > 1$ sufficiently large. This ensures that the posterior distribution of the tree complexity (number of leaf nodes) concentrates on the effective dimension $K_{total} \asymp N\epsilon_N^2$.
4. Prior support: The prior distribution on the residual variances and tree structures assigns positive mass to any Kullback-Leibler neighborhood of the true data-generating distribution.

5. Contraction rate (a consequence of items 1–4 rather than a separate assumption; Ročková and Van der Pas, 2020): the posterior distribution contracts at the optimal adaptive rate for functions with interaction order q:

$$\epsilon_N \asymp N^{-\frac{\beta}{2\beta+q}} (\log N)^{\kappa}.$$

Assumption B (Splitting-probability prior). The splitting-probability vector $\boldsymbol{s} = (s_1, \ldots, s_{P_N})$ for the prognostic functions is assigned a symmetric Dirichlet prior: $\boldsymbol{s} \sim \text{Dir}\left(\frac{\alpha_N}{P_N}, \cdots, \frac{\alpha_N}{P_N}\right)$, where the concentration parameter α_N satisfies the following two-sided rate condition. There exist constants $0 < c \leq C < \infty$ and $\gamma > 0$ such that

$$cN^{-\gamma} \leq \alpha_N \leq CN^{-\gamma} \quad \text{for all sufficiently large } N,$$

with a sufficiently large $\gamma > 0$. In shorthand, $\alpha_N \asymp N^{-\gamma}$.

Assumption C (Support condition). The true prognostic functions depend on $\boldsymbol{X}$ exactly through $\mathcal{C}^\star$: no coordinate outside $\mathcal{C}^\star$ enters f_0, and for every $X_j \in \mathcal{C}^\star$, $\inf_{f \in \mathcal{F}_{-j}} \|f_0 - f\|^2_{L_2(P)} \geq \delta > 0$, where $\mathcal{F}_{-j}$ is the class of functions not depending on X_j and δ does not depend on N or P_N. The latter holds whenever f_0 genuinely varies in X_j, and is thus a property of the data-generating process rather than an input to the method.

Model indices. For a model index $\mathcal{S}$, write $\mathcal{F}_\mathcal{S}$ for the sum-of-trees functions that split only on covariates in $\mathcal{S}$, and $m(\mathcal{D} \mid \mathcal{S})$ for the marginal likelihood of the data under this model. Under the Dirichlet prior with small α_N, each covariate that receives non-negligible splitting probability costs a prior factor of order α_N/P_N [Linero, 2018], so adding k covariates to a model index multiplies its prior probability by a factor of order $(\alpha_N/P_N)^k$; this is the only property of $\pi(\mathcal{S})$ used below.

Theorem 6.1 (**Posterior Consistency of Confounder Selection).** *Let $\mathcal{C}^\star$ be the covariate set defined above, and let $\mathcal{S}$ be the model index, i.e., the subset of covariates on which the trees may split. Under the regularity and support conditions of Appendix B (Assumptions A–C), if the concentration parameter of the sparsity-inducing Dirichlet prior satisfies $\alpha_N \asymp N^{-\gamma}$ for a sufficiently large $\gamma > 0$, then as sample size $N \to \infty$, the posterior probability of selecting the true set $\mathcal{C}^\star$ converges to 1:*

$$P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1.$$

Proof. Let $\mathcal{S}$ be any candidate model index as defined above. Throughout, the likelihood is the product of the exposure, mediator, and outcome likelihoods, with the exposure model entering through its latent Gaussian representation [Albert and Chib, 1993], so the Gaussian arguments below apply to all three components; since $\mathcal{C}^\star = \text{pa}(M) \cup \text{pa}(Y)$, the approximation error in Case 1 arises from the mediator and outcome components. We analyze the posterior odds ratio of $\mathcal{S}$ relative to $\mathcal{C}^\star$:

$$\frac{P(\mathcal{S} \mid \mathcal{D})}{P(\mathcal{C}^\star \mid \mathcal{D})} = \frac{\pi(\mathcal{S})}{\pi(\mathcal{C}^\star)} \times \frac{m(\mathcal{D} \mid \mathcal{S})}{m(\mathcal{D} \mid \mathcal{C}^\star)}.$$

We consider two cases based on the relationship between $\mathcal{S}$ and $\mathcal{C}^\star$.

Case 1: Underfitting models ($\mathcal{C}^\star \not\subseteq \mathcal{S}$).
Consider a set $\mathcal{S}$ that omits at least one true confounder $X_j \in \mathcal{C}^\star$. Since $\mathcal{F}_\mathcal{S} \subseteq \mathcal{F}_{-j}$, Assumption C yields a strictly positive approximation error. Under the assumption of Gaussian noise, the Kullback-Leibler divergence is proportional to the squared $L_2(P)$ norm between the functions:

$$KL(P_{f_0} || P_{f_\mathcal{S}}) \asymp \inf_{f \in \mathcal{F}_\mathcal{S}} \|f_0 - f\|^2_{L_2(P)} \geq \delta > 0.$$

Thus, by the general posterior-contraction theory of Ghosal et al. [2000], whose entropy and prior-mass conditions are supplied by Assumptions A.3 and A.4, the marginal likelihood ratio decays exponentially in N for some constant $c > 0$:

$$\frac{m(\mathcal{D} \mid \mathcal{S})}{m(\mathcal{D} \mid \mathcal{C}^\star)} \lesssim \exp\left\{-cN \inf_{f \in \mathcal{F}_\mathcal{S}} \|f_0 - f\|^2_{L_2(P)}\right\} \leq \exp\{-cN\delta\}.$$

This exponential decay dominates any polynomial prior term, driving the ratio to zero.

Case 2: Overfitting models ($\mathcal{C}^\star \subset \mathcal{S}$).
Consider a set $\mathcal{S}$ that includes $\mathcal{C}^\star$ plus k irrelevant variables. Since $\mathcal{C}^\star \subset \mathcal{S}$, f_0 resides in the model

space of $\mathcal{S}$, so the approximation bias is zero. The asymptotic behavior relies on the structural regularization provided by the tree prior defined in Assumption A.3.

Even though fitting noise with a larger tree could theoretically increase the likelihood by $\exp(C_1 K)$, the tree structure prior imposes an exponential penalty $\pi(K) \propto \exp(-C_2 K)$ with $C_2 > C_1$, which is ensured by taking λ sufficiently large in Assumption A.3. Due to this exponential cancellation, the marginal likelihood does not explode. Instead of growing infinitely, the posterior complexity of $\mathcal{S}$ concentrates on the same effective dimension as the true model $\mathcal{C}^\star$, i.e., $K_{\mathcal{S}} \approx K_{\mathcal{C}^\star} \asymp K_{opt}$ [Ročková and Van der Pas, 2020]. Since both models converge to the same structural complexity K_{opt}, their marginal likelihoods are asymptotically of the same order. Consequently, the ratio remains stochastically bounded:

$$\frac{m(\mathcal{D} \mid \mathcal{S})}{m(\mathcal{D} \mid \mathcal{C}^\star)} = O_p(1).$$

However, we must control for the multiplicity of such overfitting models. The number of subsets of size $|\mathcal{C}^\star| + k$ is $\binom{P-|\mathcal{C}^\star|}{k} \asymp P^k$. To ensure uniform consistency, the prior penalty must dominate this combinatorial factor. Using the sparsity-inducing Dirichlet prior with α_N (Assumption B), the prior ratio behaves as:

$$\frac{\pi(\mathcal{S})}{\pi(\mathcal{C}^\star)} \asymp \left(\frac{\alpha_N}{P}\right)^k \asymp N^{-\gamma k} P^{-k}.$$

Combining the bounded likelihood ratio and the prior penalty, the sum of posterior probabilities over all overfitting models of size k is:

$$\sum_{\mathcal{S}: \mathcal{C}^\star \subset \mathcal{S}, |\mathcal{S} \setminus \mathcal{C}^\star| = k} \frac{P(\mathcal{S} \mid \mathcal{D})}{P(\mathcal{C}^\star \mid \mathcal{D})} \asymp P^k \times N^{-\gamma k} P^{-k} \times O_p(1) = O_p(N^{-\gamma k}).$$

This vanishes as $N \to \infty$, provided that $\gamma > 0$ is chosen sufficiently large to handle any residual logarithmic growth in the likelihood ratio.

Thus, since the posterior odds ratios for all $\mathcal{S} \neq \mathcal{C}^\star$ vanish (exponentially for underfitting, polynomially for overfitting), the posterior probability of the true set $\mathcal{C}^\star$ converges to 1. □

Corollary 6.2 (High-Dimensional Consistency). *The consistency result in Theorem 6.1 extends to the high-dimensional setting where the number of covariates P_N grows with the sample size N (potentially $P_N \gg N$), provided that the growth rate satisfies $\log P_N = o(N)$. If the Dirichlet prior satisfies Assumption B with P replaced by P_N, so that each additional active covariate costs a prior factor of order $N^{-\gamma} P_N^{-1}$, then the posterior consistency holds:*

$$P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1,$$

for a sufficiently large $\gamma > 0$.

Proof. We analyze the posterior behavior under the regime where P_N grows with N. Consider first the underfitting case ($\mathcal{C}^\star \not\subseteq \mathcal{S}$). By Assumption C, every f in $\bigcup_{\mathcal{S} \not\supseteq \mathcal{C}^\star} \mathcal{F}_{\mathcal{S}}$ satisfies $\|f - f_0\|^2_{L_2(P)} \geq \delta$, so the Case 1 argument of Theorem 6.1 applies uniformly over this union (its entropy is $o(N)$ when $\log P_N = o(N)$), giving $m(\mathcal{D} \mid \mathcal{S})/m(\mathcal{D} \mid \mathcal{C}^\star) \lesssim \exp(-cN\delta)$ for all such $\mathcal{S}$ simultaneously. Since $\sum_{\mathcal{S}} \pi(\mathcal{S}) \leq 1$ and $\pi(\mathcal{C}^\star) \asymp (N^{-\gamma} P_N^{-1})^{|\mathcal{C}^\star|}$,

$$\sum_{\mathcal{S} \not\supseteq \mathcal{C}^\star} \frac{P(\mathcal{S} \mid \mathcal{D})}{P(\mathcal{C}^\star \mid \mathcal{D})} \lesssim \frac{\exp(-cN\delta)}{\pi(\mathcal{C}^\star)} \asymp \exp\big\{ -cN\delta + |\mathcal{C}^\star|(\log P_N + \gamma \log N) \big\} \to 0,$$

since $\log P_N = o(N)$; this is where the growth condition is used.

Consider next the overfitting case ($\mathcal{C}^\star \subset \mathcal{S}$) with k additional variables. The number of such models increases with the dimension as:

$$\binom{P_N - |\mathcal{C}^\star|}{k} \asymp P_N^k.$$

However, the prior probability ratio is scaled to counteract this multiplicity:

$$\frac{\pi(\mathcal{S})}{\pi(\mathcal{C}^\star)} \asymp \left(\frac{N^{-\gamma}}{P_N}\right)^k = N^{-\gamma k} P_N^{-k}.$$

Regarding the marginal likelihood ratio, the high-dimensional setting increases the metric entropy of the parameter space. Consequently, the minimax contraction rate incurs a logarithmic penalty:

$$\epsilon_N \asymp \left(\frac{N}{\log P_N}\right)^{-\frac{\beta}{2\beta+q}}.$$

Accordingly, the posterior complexity of the tree ensemble concentrates on an adjusted effective dimension:

$$K_{opt} \asymp N\epsilon_N^2 \asymp N\left(\frac{N}{\log P_N}\right)^{-\frac{2\beta}{2\beta+q}}.$$

Since the effective model size K_{opt} now implicitly depends on $\log P_N$, the marginal likelihood ratio is not strictly $O_p(1)$ but is stochastically bounded by a term growing polylogarithmically with the dimension:

$$\frac{m(\mathcal{D} \mid \mathcal{S})}{m(\mathcal{D} \mid \mathcal{C}^\star)} = O_p((\log P_N)^{c'}),$$

for some constant $c' > 0$.

Combining these factors, the sum of posterior odds for all overfitting models behaves as:

$$\sum_{\mathcal{S}} \frac{P(\mathcal{S} \mid \mathcal{D})}{P(\mathcal{C}^\star \mid \mathcal{D})} \asymp \underbrace{P_N^k}_{\text{Combinatorics}} \times \underbrace{N^{-\gamma k} P_N^{-k}}_{\text{Prior}} \times \underbrace{O_p((\log P_N)^{c'})}_{\text{Likelihood}} = O_p(N^{-\gamma k}(\log P_N)^{c'}).$$

Crucially, the combinatorial factor P_N^k is exactly cancelled by the prior penalty P_N^{-k}. The remaining term is dominated by the polynomial decay $N^{-\gamma k}$. Since $\log P_N = o(N)$ gives $(\log P_N)^{c'} \leq N^{c'}$ for large N, the sum is $O_p(N^{c'-\gamma k}) \to 0$ once $\gamma k > c'$, which holds for γ sufficiently large; the posterior probability of overfitting therefore vanishes as $N \to \infty$, and together with the underfitting bound this gives $P(\mathcal{S} = \mathcal{C}^\star \mid \mathcal{D}) \xrightarrow{P} 1$. □

Theorem 6.3 (Posterior Consistency of Natural Direct and Indirect Effects). *Let $\Psi_0 = (\delta_0, \zeta_0)$ denote the true natural direct and indirect effects, and let $\Psi = (\delta, \zeta)$ denote the random variables corresponding to these effects derived from the posterior distribution. Under Assumptions A–C (Appendix B) and the result of Theorem 6.1, the posterior distribution of the estimated effects concentrates around the true values. That is, for any $\xi > 0$,*

$$P(\|\Psi - \Psi_0\|_2 > \xi \mid \mathcal{D}) \xrightarrow{P} 0 \quad \text{as } N \to \infty,$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm for vectors.

Proof. Under sequential ignorability, the NDE and NIE are identified as:

$$\delta_0 = \int\int [E(Y \mid A=1, M=m, \mathbf{X}=\mathbf{x}) - E(Y \mid A=0, M=m, \mathbf{X}=\mathbf{x})]\, dF_{M|A=0,\mathbf{X}=\mathbf{x}}(m)\, dP(\mathbf{x}),$$
$$\zeta_0 = \int\int E(Y \mid A=1, M=m, \mathbf{X}=\mathbf{x})[dF_{M|A=1,\mathbf{X}=\mathbf{x}}(m) - dF_{M|A=0,\mathbf{X}=\mathbf{x}}(m)]\, dP(\mathbf{x}).$$

Under the BCMF parametrisation, these simplify to

$$\delta_0 = \int \tau_Y(\mathbf{x})\, dP(\mathbf{x}), \qquad \zeta_0 = \int \tau'_Y(\mathbf{x})\, \tau_M(\mathbf{x})\, dP(\mathbf{x}),$$

where $\tau_Y(\mathbf{x}) = E(Y \mid A=1, \mathbf{X}=\mathbf{x}) - E(Y \mid A=0, \mathbf{X}=\mathbf{x})$ and $\tau_M(\mathbf{x}), \tau'_Y(\mathbf{x})$ are the conditional treatment effect on M and the marginal effect of M on Y, respectively. These follow by substituting $E(Y \mid A, M, \mathbf{X}) = \mu_Y + \tau_Y A + \tau'_Y M$ into the identification formulas and using $\int m\, dF_{M|A=a,\mathbf{X}=\mathbf{x}}(m) = \mu_M(\mathbf{x}) + \tau_M(\mathbf{x})\, a$, after which the μ terms cancel in each contrast.

Also,

$$\begin{aligned} P(\|\Psi-\Psi_0\|_2>\xi \mid \mathcal{D}) &= P(\|\Psi-\Psi_0\|_2>\xi, \mathcal{S}=\mathcal{C}^\star \mid \mathcal{D}) \\ &\quad + P(\|\Psi-\Psi_0\|_2>\xi, \mathcal{S}\neq\mathcal{C}^\star \mid \mathcal{D}) \\ &\le P(\|\Psi-\Psi_0\|_2>\xi \mid \mathcal{S}=\mathcal{C}^\star, \mathcal{D}) \cdot \underbrace{P(\mathcal{S}=\mathcal{C}^\star \mid \mathcal{D})}_{\le 1} \\ &\quad + \underbrace{P(\mathcal{S}\neq\mathcal{C}^\star \mid \mathcal{D})}_{\to 0 \text{ by Thm. 6.1}} \\ &\le P(\|\Psi-\Psi_0\|_2>\xi \mid \mathcal{S}=\mathcal{C}^\star, \mathcal{D}) + o_p(1). \end{aligned} \tag{7}$$

It remains to show $P(\|\Psi-\Psi_0\|_2>\xi \mid \mathcal{S}=\mathcal{C}^\star,\mathcal{D})\to 0$. Below, hatted quantities such as $\hat\tau_M$ denote a draw from the posterior given $\mathcal{S}=\mathcal{C}^\star$; every bound holds for each draw on an event of posterior probability tending to one, so the pointwise statements transfer to the posterior probabilities in the theorem. Conditional on $\mathcal{S}=\mathcal{C}^\star$, the BART posterior contracts to the true functions $(f_{M,0}, f_{Y,0})$ at rate ε_N (Assumption A):

$$\|\hat f_M - f_{M,0}\|_{L_2(P)}\to 0, \quad \|\hat\tau_Y - \tau_{Y,0}\|_{L_2(P)}\to 0, \quad \|\hat\tau'_Y - \tau'_{Y,0}\|_{L_2(P)}\to 0, \quad \|\hat\tau_M - \tau_{M,0}\|_{L_2(P)}\to 0$$

in $P_{f_0}^N$-probability. Contraction of the components follows from that of the full regression functions, since each component is a treatment or mediator contrast, e.g. $\tau_M(\mathbf{x}) = f_M(1,\mathbf{x}) - f_M(0,\mathbf{x})$, and positivity bounds the $L_2(P)$ error of each arm by a constant multiple of the joint error. We now bound the errors in δ and ζ separately.

Error in NDE (δ). δ is a *linear* functional of τ_Y:

$$|\hat\delta - \delta_0| = \left| \int (\hat\tau_Y(\mathbf{x}) - \tau_{Y,0}(\mathbf{x}))\, dP(\mathbf{x}) \right|.$$

Applying the Cauchy–Schwarz inequality in $L_2(P)$, with $u = \hat\tau_Y - \tau_{Y,0}$ and $v \equiv 1$:

$$|\hat\delta - \delta_0| = |\langle \hat\tau_Y - \tau_{Y,0}, 1\rangle_{L_2(P)}| \le \|\hat\tau_Y - \tau_{Y,0}\|_{L_2(P)} \cdot \underbrace{\|1\|_{L_2(P)}}_{=1} = \|\hat\tau_Y - \tau_{Y,0}\|_{L_2(P)}.$$

Since $\|\hat\tau_Y - \tau_{Y,0}\|_{L_2(P)}\to 0$, $|\hat\delta - \delta_0|\to 0$.

Error in NIE (ζ). ζ is a *bilinear* functional of (τ'_Y, τ_M):

$$\hat\zeta - \zeta_0 = \int \hat\tau'_Y(\mathbf{x})\hat\tau_M(\mathbf{x})\, dP(\mathbf{x}) - \int \tau'_{Y,0}(\mathbf{x})\tau_{M,0}(\mathbf{x})\, dP(\mathbf{x}).$$

Add and subtract $\int \tau'_{Y,0}\hat\tau_M\, dP$:

$$\hat\zeta - \zeta_0 = \int (\hat\tau'_Y - \tau'_{Y,0})\hat\tau_M\, dP + \int \tau'_{Y,0}(\hat\tau_M - \tau_{M,0})\, dP.$$

For bounding the first term, by Cauchy-Schwarz:

$$\left| \int (\hat\tau'_Y - \tau'_{Y,0})\hat\tau_M\, dP \right| \le \|\hat\tau'_Y - \tau'_{Y,0}\|_{L_2(P)} \cdot \|\hat\tau_M\|_{L_2(P)}.$$

$$\|\hat\tau_M\|_{L_2(P)} \le \underbrace{\|\hat\tau_M - \tau_{M,0}\|_{L_2(P)}}_{=o_p(1) \text{ by Assumption A}} + \underbrace{\|\tau_{M,0}\|_{L_2(P)}}_{\le B_M \text{ (deterministic)}},$$

where the deterministic bound on $\|\tau_{M,0}\|_{L_2(P)}$ follows from $\tau_{M,0}\in\mathcal{H}^\beta([0,1]^{s^\star}, L)$, $s^\star = |\mathcal{C}^\star|$ (the class of Assumption A.1 restricted to the coordinates in $\mathcal{C}^\star$, on which f_0 depends by Assumption C), being bounded on a compact domain, giving $\|\tau_{M,0}\|_\infty \le B_M$ and hence $\|\tau_{M,0}\|_{L_2(P)} \le \|\tau_{M,0}\|_\infty \cdot [P(\mathcal{X})]^{1/2} = B_M$. Therefore $\|\hat\tau_M\|_{L_2(P)} = B_M + o_p(1) = O_p(1)$, and

$$\left| \int (\hat\tau'_Y - \tau'_{Y,0})\hat\tau_M\, dP \right| \le \big(B_M + o_p(1)\big)\, \|\hat\tau'_Y - \tau'_{Y,0}\|_{L_2(P)} = O_p(\varepsilon_N) = o_p(1).$$

For bounding the second term, by Cauchy–Schwarz:

$$\left| \int \tau'_{Y,0}(\hat{\tau}_M - \tau_{M,0})\, dP \right| \leq \|\tau'_{Y,0}\|_{L_2(P)} \cdot \|\hat{\tau}_M - \tau_{M,0}\|_{L_2(P)}.$$

The true function $\tau'_{Y,0} \in \mathcal{H}^\beta$ is bounded on $[0,1]^{s^\star}$: $|\tau'_{Y,0}(\mathbf{x})| \leq B'_Y$ for a deterministic constant B'_Y (Hölder functions on compact domains are bounded). Hence $\|\tau'_{Y,0}\|_{L_2(P)} \leq B'_Y < \infty$, and

$$\left| \int \tau'_{Y,0}(\hat{\tau}_M - \tau_{M,0})\, dP \right| \leq B'_Y \, \|\hat{\tau}_M - \tau_{M,0}\|_{L_2(P)} \to 0.$$

Therefore, combining the two bounds,

$$|\hat{\zeta} - \zeta_0| \leq \big(B_M + o_p(1)\big) \, \|\hat{\tau}'_Y - \tau'_{Y,0}\|_{L_2(P)} + B'_Y \, \|\hat{\tau}_M - \tau_{M,0}\|_{L_2(P)} = o_p(1).$$

The joint bound can be obtained with the previous results:

$$\|\hat{\Psi} - \Psi_0\|_2^2 = |\hat{\delta} - \delta_0|^2 + |\hat{\zeta} - \zeta_0|^2 \leq \|\hat{\tau}_Y - \tau_{Y,0}\|_{L_2(P)}^2 + \big[B_M \|\hat{\tau}'_Y - \tau'_{Y,0}\|_{L_2(P)} + B'_Y \|\hat{\tau}_M - \tau_{M,0}\|_{L_2(P)}\big]^2.$$

Each term on the right converges to 0 in probability by Assumption A, so $\|\hat{\Psi} - \Psi_0\|_2 \to 0$ in probability conditionally on $\mathcal{S} = \mathcal{C}^\star$.

Formally, for any $\xi > 0$, by the joint bound condition,

$$P(\|\Psi - \Psi_0\|_2 > \xi \mid \mathcal{S} = \mathcal{C}^\star, \mathcal{D}) \to 0,$$

and inserting back into (7):

$$P(\|\Psi - \Psi_0\|_2 > \xi \mid \mathcal{D}) \to 0 \quad \text{in } P_{f_0}^N\text{-probability.}$$

□

Appendix C: Prior and Parameter Setting for the Bayesian Causal Mediation Forest

The priors and parameter settings for the Bayesian causal mediation forest used in the main text are as follows:

- Number of trees (T): We used 200 trees for the prognostic functions and 20 trees for the modifier functions as default settings.
- Tree depth prior: We employed the prior $\alpha(1+d)^{-\beta}$, where $\alpha \in (0,1)$ and $\beta \in [0,\infty)$, for the depth d of each tree. For prognostic functions, we set $\alpha = 0.95$ and $\beta = 2$; for modifier functions, we used $\alpha = 0.5$ and $\beta = 2$.
- Residual variance prior: The residual variance σ^2 for each model followed the prior $\sigma^2 \sim \nu\lambda/\chi^2_\nu$, with $\nu = 3$. The scale parameter λ was chosen such that $P(\sigma < \hat{\sigma}) = 0.90$, where $\hat{\sigma}$ is a data-driven estimate.
- Tree alteration steps: While the original BART algorithm [Chipman et al., 2010] includes four tree alteration steps (grow, prune, change, and swap), we omit the swap step based on prior findings that sufficient mixing is achieved without it [Kapelner and Bleich, 2016]. Hence, trees are updated using only the grow, prune, and change operations.

All other priors and parameters follow the default specifications provided in Chipman et al. [2010].

BCMF model specification. Under Models (3), (4), and (5), the BCMF model assigns independent discrete uniform priors over selection probabilities for each prognostic function:

$$\begin{aligned}
\boldsymbol{s}^A &= \big(s_1^A, s_2^A, \ldots, s_P^A\big) \sim \text{discrete uniform}(1, P), \\
\boldsymbol{s}^M &= \big(s_1^M, s_2^M, \ldots, s_P^M, s_\pi^M\big) \sim \text{discrete uniform}(1, P+1), \\
\boldsymbol{s}^Y &= (s_1^Y, s_2^Y, \ldots, s_P^Y, s_\pi^Y, s_{m0}, s_{m1}) \sim \text{discrete uniform}(1, P+3).
\end{aligned}$$

DCMF model specification. Under Models (3), (4), and (5), the DCMF model assigns independent sparsity-inducing Dirichlet priors over selection probabilities for each prognostic function:

$$\begin{aligned}
\boldsymbol{s}^A &= \left(s_1^A, s_2^A, \ldots, s_P^A\right) \sim \mathcal{D}\left(\frac{1}{P}, \cdots, \frac{1}{P}\right), \\
\boldsymbol{s}^M &= \left(s_1^M, s_2^M, \ldots, s_P^M, s_\pi^M\right) \sim \mathcal{D}\left(\frac{1}{P+1}, \cdots, \frac{1}{P+1}\right), \\
\boldsymbol{s}^Y &= \left(s_1^Y, s_2^Y, \ldots, s_P^Y, s_\pi^Y, s_{m0}, s_{m1}\right) \sim \mathcal{D}\left(\frac{1}{P+3}, \cdots, \frac{1}{P+3}\right).
\end{aligned}$$

LSEM model specification. This model adopts the classical linear structural equation framework:

$$\begin{aligned}
Y_i(a,m) &= \beta_{0Y} + \boldsymbol{X}_i^\top \beta_Y + \gamma_{0Y} a + \zeta_{0Y} m + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_Y^2), \\
M_i(a) &= \beta_{0M} + \boldsymbol{X}_i^\top \beta_M + \gamma_{0M} a + \nu_i, \quad \nu_i \sim \mathcal{N}(0, \sigma_M^2),
\end{aligned}$$

where $\boldsymbol{X}_i$ denotes the vector of covariates included for confounder adjustment.

Appendix D: Pseudo-algorithm

Algorithm 1 Posterior Computation Algorithm

Require: Posterior samples from the previous iteration:

$$\boldsymbol{\phi}^{\star} = \left[\{(\mathcal{T}_t^f, \mathcal{M}_t^f) \mid t = 1, \ldots, T_f;\ f \in \{\mu, \mu_M, \tau_M, \mu_Y, \tau_Y, \tau_Y'\}\},\ \sigma_Y^2,\ \sigma_M^2,\ \boldsymbol{M}^{\text{mis}}\right]$$

for $r = 1, \ldots, R$ **do** ▷ Repeat for R MCMC iterations
  for $f \in \{\mu, \mu_M, \tau_M, \mu_Y, \tau_Y, \tau_Y'\}$ **do**
    for $t = 1, \ldots, T_f$ **do**
      for $i = 1, \ldots, n$ **do**
        Compute residual for observation i in tree t for function f: $R_{i,-t}^{f,(r)}$
      end for
      Sample tree structure: $\mathcal{T}_t^{f,(r)} \sim p\left(\mathcal{T}_t^f \mid R_{1,-t}^{f,(r)}, \ldots, R_{n,-t}^{f,(r)}, \sigma_f^2\right)$
      Sample terminal node parameters: $\mathcal{M}_t^{f,(r)} \sim p\left(\mathcal{M}_t^f \mid \mathcal{T}_t^{f,(r)}, R_{1,-t}^{f,(r)}, \ldots, R_{n,-t}^{f,(r)}, \sigma_f^2\right)$
      where $\sigma_f^2 = \begin{cases} 1, & \text{if } f = \mu \\ \sigma_M^2, & \text{if } f \in \{\mu_M, \tau_M\} \\ \sigma_Y^2, & \text{otherwise} \end{cases}$
    end for
  end for
  Sample residual variances:

$$(\sigma_M^2)^{(r)} \sim \text{Inv-Gamma}\left(a_{\sigma_M} + \frac{n}{2},\ b_{\sigma_M} + \frac{1}{2}\sum_{i=1}^{n}(M_i - \hat{M}_i)^2\right)$$

$$(\sigma_Y^2)^{(r)} \sim \text{Inv-Gamma}\left(a_{\sigma_Y} + \frac{n}{2},\ b_{\sigma_Y} + \frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right)$$

  Update $\boldsymbol{s}^{(r)}$ based on Eq. (6) and the M-H algorithm
end for

Appendix E: Simulation Setup

In seven different scenarios, $P = 100$ potential confounders (except for Scenarios 5-7) are independently generated from $Unif(0, 1)$ for $X_1 - X_2$ and from $N(0, 1)$ for $X_3 - X_{100}$ where only 5 of them ($X_1 - X_5$) are true confounders:

- Scenario 1: Heterogeneous effects with non-linear terms.

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5}) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\ 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} \\
&\quad - \exp(0.3X_{i,5}) - 1.5A_i|X_{i,5} + 0.3| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\ 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + A_iX_{i,4} \\
&\quad + \exp(0.5X_{i,5})
\end{aligned}$$

  where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

- Scenario 2: Heterogeneous effects with non-linear terms with two additional predictors in the Y model.

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5}) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\ 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} \\
&\quad - \exp(0.3X_{i,5}) - 1.5A_i|X_{i,5} + 0.3| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\ 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + 2X_{i,4} \\
&\quad + \exp(0.5X_{i,5}) - A_i|X_{i,6}| - A_i|X_{i,7} + 1|
\end{aligned}$$

- Scenario 3: Heterogeneous effects with non-linear terms with two additional instruments in the A model.

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5} + 1.5|X_{i,6}| \\
&\quad -|X_{i,7} + 1|) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} - \exp(0.3X_{i,5}) \\
&\quad -1.5A_i|X_{i,5} + 0.3| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + 2X_{i,4} \\
&\quad + \exp(0.5X_{i,5})
\end{aligned}$$

- Scenario 4: Heterogeneous effects with non-linear terms with two additional instruments in the A model and two additional instruments in M.

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5} + 1.5|X_{i,6}| \\
&\quad -|X_{i,7} + 1|) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} - \exp(0.3X_{i,5}) \\
&\quad -1.5A_i|X_{i,5} + 0.3| + 0.2|X_{i,8}| - 0.2|X_{i,9}| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + 2X_{i,4} \\
&\quad + \exp(0.5X_{i,5})
\end{aligned}$$

- Scenarios 5 & 6: Heterogeneous effects with non-linear terms in high-dimensional settings ($N < P = 500$ & $N < P = 2000$, respectively).

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5} + 1.5|X_{i,6}| \\
&\quad -|X_{i,7} + 1|) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} - \exp(0.3X_{i,5}) \\
&\quad -1.5A_i|X_{i,5} + 0.3| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + 2X_{i,4} \\
&\quad + \exp(0.5X_{i,5})
\end{aligned}$$

- Scenario 7: Heterogeneous effects with non-linear terms with $P = 20$.

$$\begin{aligned}
P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5} + 1.5|X_{i,6}| \\
&\quad -|X_{i,7} + 1|) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.1^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1.3A_i - 0.5h_1(X_{i,1}) + 0.5h_2(X_{i,2}) + |X_{i,3} + 1| + 1.5X_{i,4} - \exp(0.3X_{i,5}) \\
&\quad -1.5A_i|X_{i,5} + 0.3| \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= -A_i + h_1(X_{i,1}) + 1.5h_2(X_{i,2}) + 2|X_{i,3} + 1| - 1.5M_i + 2X_{i,4} \\
&\quad + \exp(0.5X_{i,5})
\end{aligned}$$

In all scenarios, the first two variables are included in the models through non-linear functions $h_1(x) = (-1)^{I(x<0)}$ and $h_2(x) = (-1)^{I(x\geq 0)}$ and the remaining covariates are also added either with an absolute function or with an interaction term.

Appendix F: Additional Simulations

F.1. Sensitivity Analyses

In the main manuscript, we conducted a sensitivity analysis of the prior specification by examining how performance varies with the number of trees used in the BART priors, and the results are presented in Table 2. In addition, we aim to further investigate how performance changes with respect

Table 3: Results (bias, absolute bias, mean squared error, and coverage) for additional simulation scenarios (F.1-F.4)

.

		BIAS		Abs.BIAS		MSE		Coverage	
Scenario	Setting	DE	IE	DE	IE	DE	IE	DE	IE
F.1	$\alpha = 5$	-0.023	-0.042	0.096	0.173	0.021	0.056	0.99	0.99
F.2	unmeasured confounder	0.052	-0.038	0.426	1.080	0.316	1.878	0.82	0.60
F.3	model misspecification	-0.016	0.037	0.421	0.454	0.288	0.408	0.77	0.66
F.4	complex backdoor path	-0.016	-0.003	0.059	0.130	0.008	0.032	0.99	0.99

to different hyper-parameter values.

In our framework, we place a hyper-prior on α using the transformation $\alpha/(\alpha + P) \sim Beta(0.5, 1)$. This allows α to be adaptively tuned during the MCMC process. From our experiments, we observed that the value of α/P, rather than α itself, plays a pivotal role in controlling the overall sparsity. Therefore, we believe that as long as either α becomes small or P becomes large, the desired sparsity level can be achieved.

We further investigated how sensitive the model is to a fixed α by conducting an additional experiment based on the following scenario:

$$\begin{aligned} P(A_i = 1) &= \sigma(-0.3 + X_{i,1} - 1.5X_{i,2} + 0.8|X_{i,3} - 1| + 0.5X_{i,4}X_{i,5}) \\ M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.5^2) \\ \mu_M(A_i, \boldsymbol{X}_i) &= 1 + 0.9A_i + X_{i,1} - 0.5X_{i,2} + 0.5|X_{i,3} - 1| + 0.5X_{i,4}^2 - \sin(X_{i,5}) \\ Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\ \mu_Y(A_i, M_i, \boldsymbol{X}_i) &= X_{i,1} + 1.5X_{i,2} + 0.5X_{i,3}X_{i,4} - 0.4X_{i,5}^2 - 1.5M_i + A_i \end{aligned}$$

where $X_{i,j} \sim N(0, 1)$ for $i = 1, \ldots, 250; j = 1, \ldots, P$ with $P = 100$. With the Beta prior, the posterior mean of α was 1.49, and the posterior standard deviation was 0.22. Given $P = 100$, this implies $\alpha/P \approx 0.0149$. Now, we fixed $\alpha = 5$, which results in $\alpha/P = 0.05$, and examined the impact on the results. As shown in Table 3 of Appendix F.1, the overall results did not change substantially. This confirms that the absolute value of α is less critical than maintaining a sufficiently small ratio $\alpha/P \ll 1$, which effectively induces sparsity.

F.2. Unmeasured Confounder

The assumption underlying Theorem 3.2 is likely to hold in high-dimensional settings with many observed covariates, but is more vulnerable in low-dimensional settings where the chance of missing all relevant variables on a backdoor path increases. To examine this empirically, we conducted two analyses based on Scenario 2.

Analysis 1: A single unmeasured confounder. We removed the true confounder X_5 from the observed covariate set. As shown in Table 3, the absolute bias of the indirect effect (IE) more than doubled compared to the main Scenario 2 results, MSE increased nearly fourfold, and coverage dropped substantially.

Analysis 2: Varying signal strength of the unmeasured confounder. To assess how performance degrades with the strength of unmeasured confounding, we modified Scenario 2 by introducing an unmeasured variable U with effects on A, M, and Y scaled by ρ:

$$\begin{aligned} P(A_i = 1) &= \Phi(0.5 + h_1(X_{i,1}) + h_2(X_{i,2}) - 0.5|X_{i,3} - 1| + 1.5X_{i,4}X_{i,5} + \rho U_i)\,, \\ M_i &= \mu_M(A_i, \boldsymbol{X}_i) + \rho U_i + \epsilon_i, \\ Y_i &= \mu_Y(A_i, M_i, \boldsymbol{X}_i) + \rho U_i + \eta_i, \end{aligned}$$

where $U_i \sim N(0, 1)$, μ_M and μ_Y follow the same specification as Scenario 2, and the signal strength is set to $\rho \in \{0.5,\ 1.0,\ 1.5\}$ for weak, moderate, and strong unmeasured confounding, respectively. Results are summarized in Table 4.

Table 4: Performance under varying strength of unmeasured confounding U.

	BIAS		Abs.BIAS		MSE		Coverage	
Strength	DE	IE	DE	IE	DE	IE	DE	IE
Weak ($\rho = 0.5$)	-0.03	0.04	0.47	0.52	0.35	0.51	0.75	0.56
Moderate ($\rho = 1$)	-0.05	-0.07	0.66	0.72	0.64	0.89	0.57	0.34
Strong ($\rho = 1.5$)	-0.05	-0.42	0.68	0.83	0.67	1.27	0.57	0.39

Performance degrades monotonically as the signal strength of U increases, consistent with the single-confounder result above. We emphasize that this degradation is an inherent limitation shared by all causal inference methods under unmeasured confounding-no existing method, parametric or nonparametric, can avoid it without additional assumptions. Formal sensitivity analysis for Bayesian nonparametric mediation models remains underdeveloped and is beyond the present scope; we leave this for future work. We note, however, that our framework is specifically designed for settings with a rich high-dimensional covariate pool, which reduces the likelihood that all variables on a given backdoor path go unmeasured compared to low-dimensional settings.

F.3. Model Misspecification

Given that our model is nonparametric, structural model misspecification is less of a concern. Instead, performance degradation is more likely when important variables are excluded (unmeasured). To assess this, we conducted an additional analysis based on Scenario 2, in which a predictive but non-confounding variable X_6 is missing (i.e., outcome-model misspecification).

The results in Table 3 of F.3 show no noticeable difference from those of the main Scenario 2. That is, when an additional variable is a predictor but not a confounder, its exclusion from the model (i.e., even if model misspecification occurs) does not significantly impact overall performance. In contrast, when a true confounder is missing (as discussed in the previous subsection), we observe substantial performance degradation.

F.4. Complex Backdoor Path

In this subsection, we clarify that our method employs a sparsity-inducing prior for the selection vector $\boldsymbol{s}$, which naturally favors smaller confounder sets rather than maximal ones. If X_i affects downstream variables (A, M, and/or Y) only through X_j, our method tends to select either X_j or X_i, typically the one with a stronger marginal effect. In other words, our method prefers to select direct causes of A, M, and/or Y, rather than general ancestors, when they are sufficient to block backdoor paths.

To illustrate this behavior empirically, we conducted a new experiment based on the following scenario:

$$
\begin{aligned}
P(A_i = 1) &= \sigma(-0.3 + X_{i,1} - 1.5X_{i,2} + 0.8|X_{i,3} - 1| + 0.5X_{i,4}X_{i,5}) \\
M_i &\sim N(\mu_M(A_i, \boldsymbol{X}_i),\, 0.5^2) \\
\mu_M(A_i, \boldsymbol{X}_i) &= 1 + 0.9A_i + X_{i,1} - 0.5X_{i,2} + 0.5|X_{i,9} - 1| + 0.5X_{i,10}^2 - \sin(X_{i,11}) \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= X_{i,1} + 1.5X_{i,2} + 0.5X_{i,9}X_{i,10} - 0.4X_{i,11}^2 - 1.5M_i + A_i
\end{aligned}
$$

where

- X_6 influences both X_3 and X_9,
- X_7 influences both X_4 and X_{10},
- X_8 influences both X_5 and X_{11}.

The results are provided in Table 3 of F.4 and Figure 6. Our method successfully identified X_1, X_2, X_9, X_{10} and X_{11} as confounders, which are sufficient to block the backdoor paths related to $X_6 - X_8$. This provides strong empirical support that our method identifies a sparse set of direct causes as confounders.

F.5. Dependent Covariates: the DAG of Figure 1(b)

This scenario replicates Figure 1(b) with $N = 250$ and $P = 100$. Exogenous $X_2, X_5, X_6, Z, W_M, W_Y$ and 89 noise covariates are $N(0,1)$, and dependent covariates follow $X_j = 0.7\, X_{\mathrm{pa}(j)} + N(0, 0.75^2)$ along the chains $X_2 \to X_1$, $X_2 \to X_3$, $X_5 \to X_4$, and $X_6 \to X_7 \to X_8$. The models are

$$\begin{aligned} P(A_i = 1) &= \sigma(-0.3 + X_{i,1} + 0.8X_{i,6} + 0.8Z_i), \\ M_i &= 1 + 0.9A_i + X_{i,3} + 0.8|X_{i,4} - 1| + 0.5W_{M,i} + N(0, 0.5^2), \\ Y_i &= A_i - 1.5M_i + X_{i,5} - 0.4X_{i,5}^2 + 1.5X_{i,8} + 0.5W_{Y,i} + N(0, 0.3^2), \end{aligned}$$

so that $\mathcal{C}^\star = \{X_3, X_4, X_5, X_8, W_M, W_Y\}$, with X_1, X_2, X_6, X_7 ancestors on the backdoor paths and Z a pure instrument. A stress variant sets $X_3 = 0.9X_2 + N(0, 0.45^2)$ and the coefficient of X_3 in M to 0.6, so that the ancestor X_2 is as strongly associated with M as the direct parent X_3. Across 200 replicates BCMF-CS selects exactly $\mathcal{C}^\star$, with PIPs of at least 0.999 for all six variables in all three models, while the ancestors, the instrument, and the noise covariates stay below 0.13; the direct and indirect effects have absolute bias 0.05 and 0.14, MSE 0.004 and 0.033, and coverage 1.00. In the stress variant X_3 keeps a PIP above 0.999 while X_2 stays below 0.04, with essentially identical estimation accuracy.

F.6. External Baseline: Outcome-Adaptive LASSO

We adapt the outcome-adaptive LASSO [Shortreed and Ertefaie, 2017] to mediation by applying it to each confounding pathway (A–M, A–Y, and M–Y adjusting for A) and taking the union of the three selections; the selected set is then passed to the same BCMF estimation step used by BCMF-CS, so that the comparison isolates selection. On Scenarios 3 and 5 the baseline recovers every true confounder (TPR 1.00 in all three models) but admits false positives (precision 0.63–0.72), and its estimates deteriorate accordingly: on Scenario 3 the absolute bias is 0.47/0.63 (DE/IE) with coverage 0.65/0.39, against 0.31/0.49 and 0.77/0.59 for BCMF-CS; on Scenario 5 it is 0.47/0.63 with coverage 0.65/0.40, against 0.09/0.33 and 0.99/0.80.

F.7. Interval Widths

Table 5 reports average 95% interval widths in Scenario 3, where indirect-effect coverage is lowest. BCMF-CS pairs the widest intervals among the Bayesian methods with the best coverage, so the competing methods fail by overconfidence, whereas LSEM's intervals are nearly ten times wider yet cover only 39% of the time, reflecting misspecification bias.

Table 5: Average interval widths and coverage in Scenario 3.

Method	Width (DE)	Width (IE)	Coverage (DE)	Coverage (IE)
BCMF-CS	0.32	0.22	0.77	0.59
BCMF	0.45	0.16	0.52	0.17
DCMF	0.30	0.11	0.47	0.22
LSEM	1.86	2.00	0.54	0.39

Appendix G: ACTG175 dataset

The ACTG175 dataset, accessible through the R package `speff2trial`, comprises 27 variables. Of these, 15 pretreatment variables are identified as potential confounders, with descriptions based on the package documentation:

- `age` age in years at baseline
- `wtkg` weight in kg at baseline
- `hemo` hemophilia (0=no, 1=yes)
- `homo` homosexual activity (0=no, 1=yes)
- `drugs` history of intravenous drug use (0=no, 1=yes)

- `karnof` Karnofsky score (on a scale of 0-100)
- `oprior` non-zidovudine antiretroviral therapy prior to initiation of study treatment (0=no, 1=yes)
- `z30` zidovudine use in the 30 days prior to treatment initiation (0=no, 1=yes)
- `preanti` number of days of previously received antiretroviral therapy
- `race` race (0=white, 1=non-white)
- `gender` gender (0=female, 1=male)
- `str2` antiretroviral history (0=naive, 1=experienced)
- `strat` antiretroviral history stratification (1='antiretroviral naive', 2='>1 but 52 weeks of prior antiretroviral therapy', 3='>52 weeks')
- `symptom` symptomatic indicator (0=asymptomatic, 1=symptomatic)
- `cd40` CD4 T cell count at baseline

Appendix H: Extension to Multiple Mediators

In practice, multiple mediators may exist. In such settings, researchers can isolate a mediator of primary scientific interest and interpret the indirect effect as that which is transmitted through this mediator. Effects transmitted through other paths are typically subsumed into the direct effect. This practice is also standard in the literature and provides meaningful interpretation even in complex settings.

That said, we also acknowledge the growing interest in mediation analysis involving multiple mediators. While our current theorems focus on the single-mediator case, the proposed model architecture is extensible. Specifically, building on the framework of Imai and Yamamoto [2013], we adapt our model to accommodate multiple causally independent mediators (with two mediators in this toy example) as follows:

We construct separate mediator models (Eq. 4) for both mediators, M_1 and M_2:

$$\begin{aligned} M_{1i} &= \mu_{M_1}(\boldsymbol{X}_i, \hat{\pi}_i) + \tau_{M_1}(\boldsymbol{X}_i)A_i + \epsilon_{1i} \\ M_{2i} &= \mu_{M_2}(\boldsymbol{X}_i, \hat{\pi}_i) + \tau_{M_2}(\boldsymbol{X}_i)A_i + \epsilon_{2i} \end{aligned}$$

The outcome model (Eq. 5) is expanded to include four clever covariates derived from the two mediator models. Both M_1 and M_2 are included as linear terms in the outcome regression model:

$$Y_i = \mu_Y(\boldsymbol{X}_i, \hat{\pi}_i, \hat{m}_{10i}, \hat{m}_{11i}, \hat{m}_{20i}, \hat{m}_{21i}) + \tau_Y(\boldsymbol{X}_i)A_i + \tau'_Y(\boldsymbol{X}_i)M_{1i} + \tau''_Y(\boldsymbol{X}_i)M_{2i} + \eta_i,$$

where $\hat{m}_{jai} = \hat{\mathbb{E}}(M_{ji} \mid A_i = a, \boldsymbol{X}_i)$ for $j = 1, 2$. Separate selection vectors $\boldsymbol{s}^{M_1}$ and $\boldsymbol{s}^{M_2}$ are defined for the two mediators. The common selection vector $\boldsymbol{s}$ now includes parameters for all four clever covariates, resulting in a total dimension of $P + 5$:

$$\begin{aligned} \boldsymbol{s}^A &= \left(\frac{s_1}{1 - (s_\pi + s_{m10} + s_{m11} + s_{m20} + s_{m21})}, \cdots, \frac{s_P}{1 - (s_\pi + s_{m10} + s_{m11} + s_{m20} + s_{m21})} \right) \\ \boldsymbol{s}^{M_1} &= \left(\frac{s_1}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})}, \cdots, \frac{s_P}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})}, \right. \\ &\quad \left. \frac{s_\pi}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})} \right) \\ \boldsymbol{s}^{M_2} &= \left(\frac{s_1}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})}, \cdots, \frac{s_P}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})}, \right. \\ &\quad \left. \frac{s_\pi}{1 - (s_{m10} + s_{m11} + s_{m20} + s_{m21})} \right) \\ \boldsymbol{s}^Y &= (s_1, \cdots, s_P, s_\pi, s_{m10}, s_{m11}, s_{m20}, s_{m21}). \end{aligned}$$

To empirically evaluate this extension, we implemented the model in a synthetic setting involving two causally independent mediators (with 100 potential confounders $\boldsymbol{X}_i$'s). Specifically, the data-

generating models for the two mediators and the outcome are as follows:

$$\begin{aligned}
P(A_i = 1) &= \sigma(-0.3 + X_{i,1} - 1.5X_{i,2} + 0.8|X_{i,3} - 1| + 0.5X_{i,4}X_{i,5}) \\
M_{1i} &\sim N(\mu_{M_1}(A_i, \boldsymbol{X}_i),\, 0.5^2) \\
\mu_{M_1}(A_i, \boldsymbol{X}_i) &= 1 + 1.2A_i + X_{i,1} - 0.5X_{i,2} + 0.5|X_{i,3} - 1| - 0.5X_{i,4}^2 - \sin(X_{i,5}) \\
M_{2i} &\sim N(\mu_{M_2}(A_i, \boldsymbol{X}_i),\, 0.5^2) \\
\mu_{M_2}(A_i, \boldsymbol{X}_i) &= 1 + 1.5A_i + 0.5X_{i,1} - X_{i,2} + 0.75|X_{i,3} - 1| - 0.4X_{i,4}^2 - \sin(X_{i,5}) \\
Y_i &\sim N(\mu_Y(A_i, M_i, \boldsymbol{X}_i),\, 0.3^2) \\
\mu_Y(A_i, M_i, \boldsymbol{X}_i) &= X_{i,1} + 1.5X_{i,2} + 0.5X_{i,3}X_{i,4} - 0.4X_{i,5}^2 - M_{1i} - 0.7M_{2i} + A_i.
\end{aligned}$$

The true (total) indirect and direct effects are -2.25 and 1.0, respectively. Table 6 summarizes the result. These results confirm that the proposed framework can be naturally extended to multiple-

Table 6: Results (bias, absolute bias, mean squared error, and coverage) for the scenario with multiple mediators (two mediators).

	BIAS		Abs.BIAS		MSE		Coverage	
	DE	IE	DE	IE	DE	IE	DE	IE
Multiple	0.302	-0.354	0.315	0.380	0.126	0.202	0.58	0.97

mediator settings without requiring fundamental changes to the algorithm structure. As the number of mediators increases, variable importance from individual mediator models may dominate the overall selection probability vector. To address this, we plan to extend the framework to a more general setting with multiple mediators in future work.

Appendix I: Tables and Figures

Table 7: Operating characteristics (true positive rate, false positive rate, positive predictive value, and F1 score) for different scenarios: Scenario 1 (base), Scenario 2 (with two predictors), Scenarios 3 and 4 (with two instruments in exposure and mediator models, respectively), Scenario 5 ($N < P = 500$), and Scenario 7 ($P = 20$). Selection results are reported separately for the treatment model (A), mediator model (M), and outcome model (Y).

		TPR			FPR			PPV			F1		
Scenario	Method	A	M	Y	A	M	Y	A	M	Y	A	M	Y
(1)	BCMF-CS	1	1	1	0.00	0.00	0.00	1	1	1	1	1	1
	BCMF	1	1	1	1	0.72	0.98	0.05	0.07	0.05	0.09	0.13	0.09
	DCMF	0.44	1	0.82	0.00	0.00	0.00	0.82	1	0.99	0.55	1	0.89
(2)	BCMF-CS	1	1	1	0.02	0.02	0.02	0.72	0.72	0.72	0.83	0.83	0.83
	BCMF	1	1	1	1	0.72	0.99	0.05	0.07	0.05	0.09	0.13	0.09
	DCMF	0.45	1	0.87	0.01	0.00	0.02	0.84	1	0.68	0.57	1	0.77
(3)	BCMF-CS	1	1	1	0.00	0.00	0.00	1	1	1	1	1	1
	BCMF	1	1	1	1	0.72	0.98	0.05	0.07	0.05	0.09	0.13	0.09
	DCMF	0.41	1	0.87	0.02	0.00	0.00	0.51	1	0.99	0.45	1	0.93
(4)	BCMF-CS	1	1	1	0.00	0.00	0.00	0.95	0.94	0.95	0.97	0.96	0.97
	BCMF	1	1	1	1	0.82	0.97	0.05	0.06	0.05	0.10	0.11	0.10
	DCMF	0.41	1	0.99	0.03	0.01	0.00	0.51	0.83	0.99	0.44	0.91	0.99
(5)	BCMF-CS	1	1	1	0.00	0.00	0.00	0.99	1	0.99	0.99	1	0.99
	BCMF	0.28	1	0.99	0.01	0.01	0.02	0.76	0.80	0.73	0.43	0.88	0.83
	DCMF	0.09	1	0.92	0.00	0.00	0.00	0.92	1	0.99	0.35	1	0.95
(7)	BCMF-CS	1	1	1	0.00	0.00	0.00	0.99	0.99	0.99	0.99	1	0.99
	BCMF	0.29	1	0.99	0.01	0.02	0.02	0.79	0.80	0.71	0.45	0.88	0.82
	DCMF	0.11	1	0.93	0.00	0.00	0.00	0.91	1	0.99	0.34	1	0.96

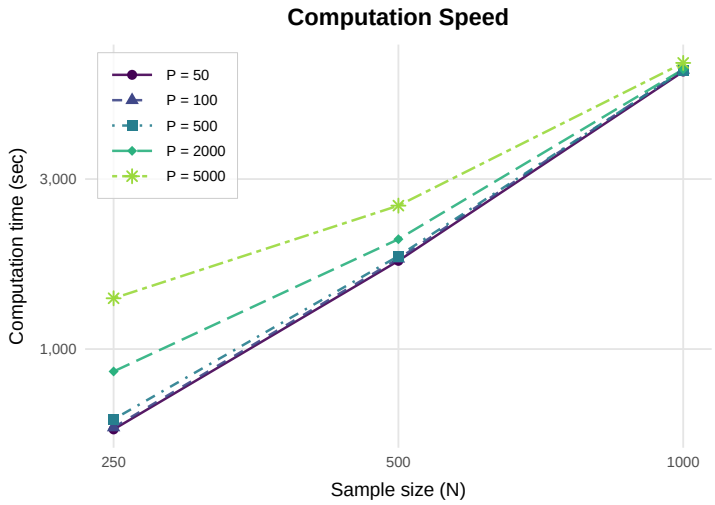

Figure 3: Scaling of computation time with the sample size N and the number of potential confounders P. The grid spans both low- and high-dimensional regimes, including settings with $P \gg N$.

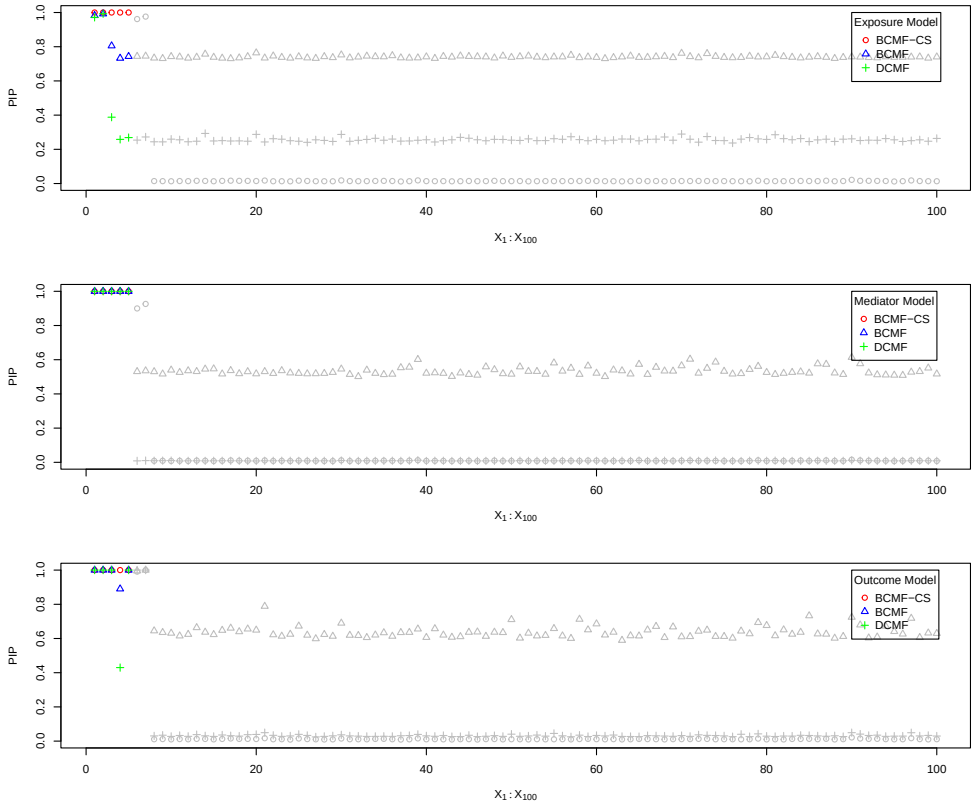

Figure 4: [Scenario 2] Posterior inclusion probabilities (PIP) from three competing methods (BCMF-CS, BCMF, and DCMF) for the exposure, mediator, and outcome models. The true confounders are represented by colored points, while the noise variables are depicted as grey points.

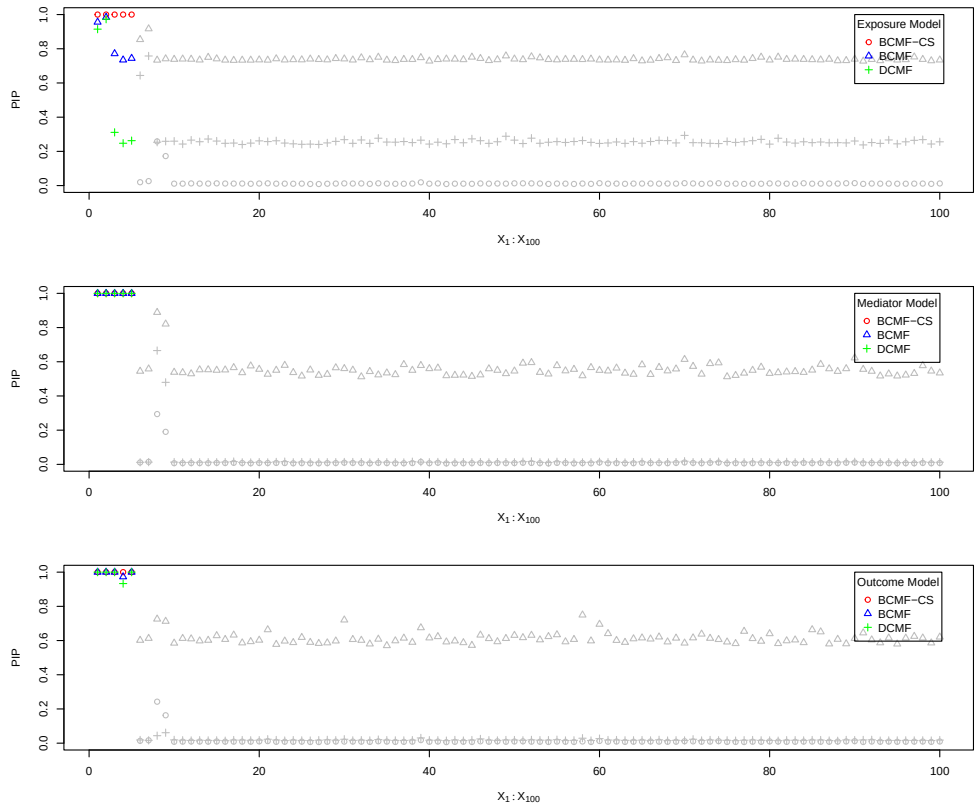

Figure 5: [Scenario 4] Posterior inclusion probabilities (PIP) from three competing methods (BCMF-CS, BCMF, and DCMF) for the exposure, mediator, and outcome models. The true confounders are represented by colored points, while the noise variables are depicted as grey points.

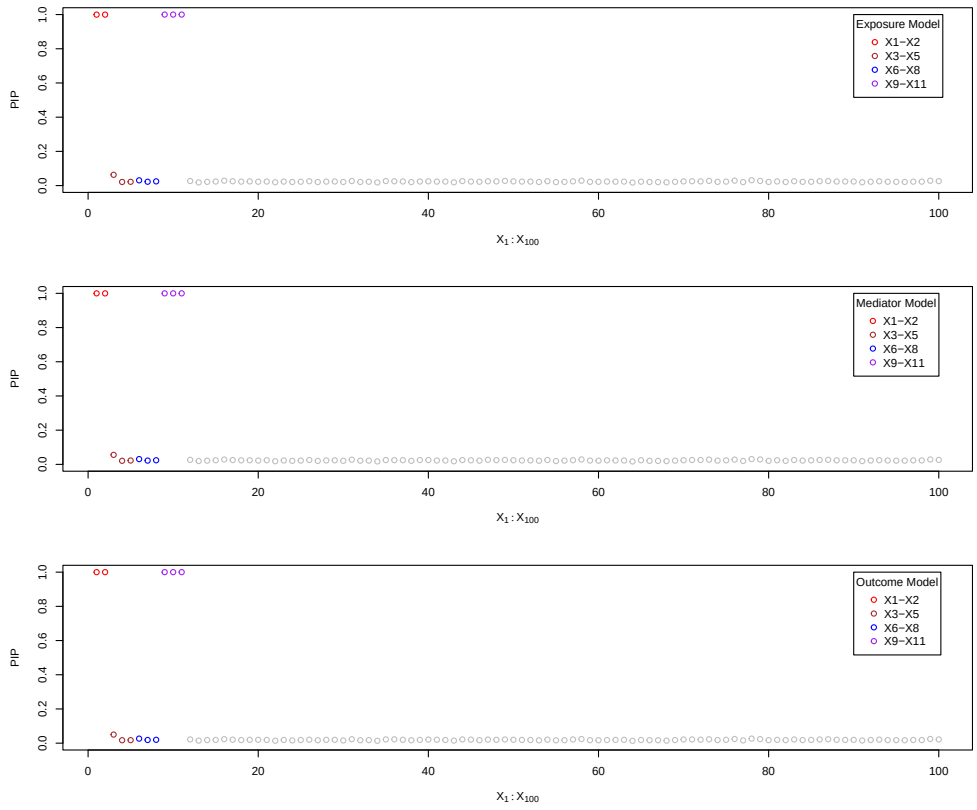

Figure 6: [Scenario F.4] Posterior inclusion probabilities (PIP) for the exposure, mediator, and outcome models under the complex backdoor path structure. The true confounders are represented by colored points, while the noise variables are depicted as grey points.

Exposure-model PIPs in the ACTG175 analysis. Because the three models share the selection vector $\boldsymbol{s}$, a covariate selected for the mediator and outcome models is also used by the exposure-model trees, which raises its exposure-model PIP regardless of its association with treatment. For cd40 and preanti that association is negligible: regressing A on the two variables gives a likelihood-ratio p-value of 0.20, and the predicted treatment probabilities range only from 0.66 to 0.82, against an overall treatment rate of 0.76.

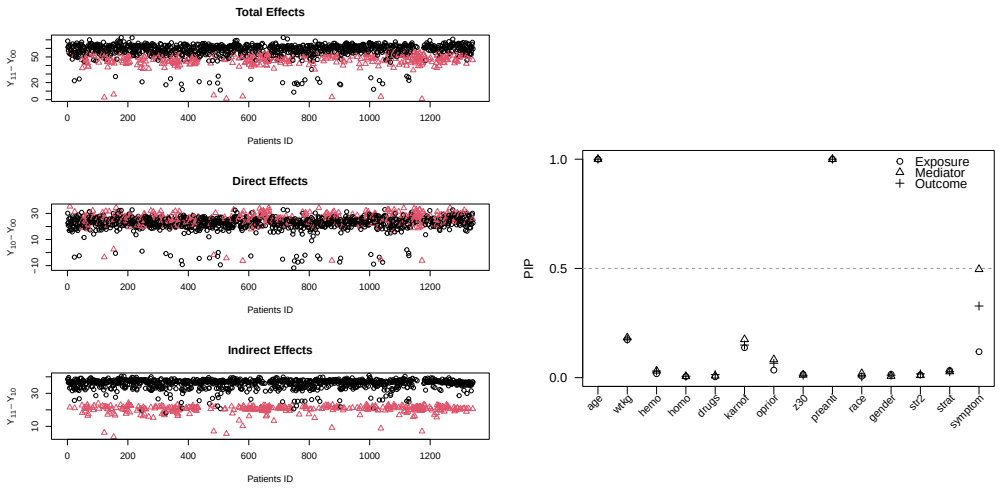

Figure 7: Results based on the ACTG175 dataset after removing "cd40", which was identified as a confounder in the main analysis. (Left) Estimated individual total, direct, and indirect effects, ordered by patient ID; red triangles represent non-white patients and black circles represent white patients. (Right) Posterior inclusion probabilities of potential confounders in the exposure, mediator, and outcome models. Notably, "age"-a parent of "cd40" in the causal graph-is now selected, indicating that the proposed method substitutes an ancestor when a direct cause is removed. The effect estimates remain largely unchanged from the main analysis.

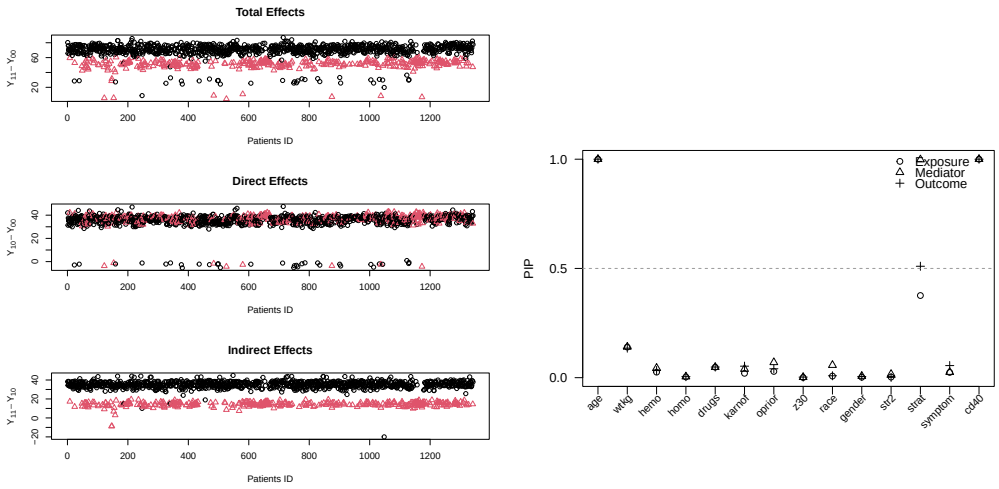

Figure 8: Results based on the ACTG175 dataset after removing "preanti", which was identified as a confounder in the main analysis. (Left) Estimated individual total, direct, and indirect effects, ordered by patient ID; red triangles represent non-white patients and black circles represent white patients. (Right) Posterior inclusion probabilities of potential confounders in the exposure, mediator, and outcome models. Notably, "age"-a parent of "preanti" in the causal graph-is now selected, indicating that the proposed method substitutes an ancestor when a direct cause is removed. The effect estimates remain largely unchanged from the main analysis.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The proposed method effectively selects confounders relevant to mediation analysis while flexibly estimating heterogeneous effects, as clearly stated in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The key assumption required for the mediation disjunctive cause criterion that all variables on backdoor paths are included in the potential confounder set is clearly stated as a limitation.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The proofs of Theorems 1, 2, and 3 are all thoroughly demonstrated in the Appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: To ensure transparency and reproducibility of the results, we provide the computational algorithm, detailed prior specifications, and parameter settings in the Appendix. Additionally, code for replicating the simulation and application results is included in the supplementary material.

Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: To ensure the reproducibility of the results, the simulation and application code is provided in the supplementary material.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All settings, including the (hyper)parameters used for the experiments, are described in the main manuscript and the supplementary material.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Statistical uncertainty is reflected in Table 1 through the mean squared error and coverage metrics.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Computation time is reported in the Appendix, with a detailed explanation provided in the main manuscript.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Yes, this paper conform the the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

Justification: This paper presents foundational research and does not pose any direct negative societal impacts.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: No such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All data and methods are explicitly mentioned.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: All data, code, and documentation are given in the Supplementary Material.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve human subject.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: This paper does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: This paper does not involve LLMs any important components.

    Guidelines:

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