Optimal Experiments for Partial Causal Effect Identification

Tobias Maringgele
Technical University of Munich
tobias.maringgele@tum.de

Jalal Etesami
Technical University of Munich
j.etesami@tum.de

Abstract

Causal queries are often only partially identifiable from observational data, and experiments that could tighten the resulting bounds are typically costly. We study the problem of selecting, prior to observing experimental outcomes, a cost-constrained subset of experiments that maximally tightens bounds on a target query. We formalize this as the *max-potency problem*, where epistemic potency measures the worst-case reduction in bound width guaranteed by an experiment, and show that this problem is NP-hard via a reduction from 0-1 knapsack. Building on the polynomial-programming framework of Duarte et al. [2024], we give a general procedure for evaluating epistemic potency in discrete settings. To control the super-exponential search space, we introduce two graphical pruning criteria that depend only on the causal graph and the query: a novel path-interception rule that exploits district structure to certify zero potency in linear time, and an identifiability check based on the ID algorithm. On Erdős–Rényi random graphs and 11 *bnlearn* benchmark networks, the two criteria together prune 50–88% of candidate experiments on average without solving a single polynomial program. For the general subset search, we show that ID-pruned experiments are combinatorially inert, yielding a super-exponential reduction in the number of subsets evaluated. We close with an end-to-end demonstration on observational NHANES data, selecting optimal experiments for estimating the effect of physical activity on diabetes.

1 Introduction

A central goal of causal inference is to estimate the effect of a treatment on an outcome. Population-level summaries such as the average treatment effect (ATE) capture aggregate behavior, but many decisions hinge on finer-grained counterfactual quantities—for example, the probability that a treatment is both necessary and sufficient to produce an outcome (PNS) [Pearl, 1999, Mueller and Pearl, 2023]. Such queries are rarely point-identifiable from observational data; the best one can do is derive *bounds* [Balke and Pearl, 1997, Tian and Pearl, 2000], which are often too wide to be decision-relevant. Tightening them requires additional data, and additional data means experiments, which are costly and subject to budget constraints.

This motivates the central question of this paper: *given a causal graph, an observational distribution, and a budget, which subset of experiments, chosen ex ante, most tightens bounds on a target query?* Answering it well means reasoning about how informative an experiment will be *before* running it, unlike sequential designs that adapt to realized outcomes [Ailer et al., 2024], and doing so over a search space that grows super-exponentially in the number of variables.

Our contributions are as follows:

- We formalize this as the *max-potency problem*, where epistemic potency measures the worst-case reduction in bound width an experiment guarantees, and show it is NP-hard via a reduction from 0-1 knapsack (Section 4).

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

- We give a general procedure for evaluating epistemic potency in discrete settings, building on the polynomial-programming framework of Duarte et al. [2024] (Section 5).
- We introduce two graphical pruning criteria that depend only on the causal graph and the query to certify zero potency of candidate experiments without solving any polynomial program: a novel path-interception rule that runs in linear time by exploiting the graph's district structure, and an identifiability check based on the ID algorithm [Shpitser and Pearl, 2006]. We further show that ID-pruned experiments are *combinatorially inert*: they can be removed from the subset search entirely (Section 6).
- On Erdős–Rényi random graphs and 11 *bnlearn* benchmark networks, the two criteria together prune 50–88% of candidate experiments on average when selecting a single optimal experiment. For the general subset search, ID-inertness alone reduces the search space exponent by roughly 60% on average—a super-exponential reduction in the number of subsets that need to be evaluated. We demonstrate the full pipeline end-to-end on observational NHANES data (Section 7).

1.1 Related Work

Our work bridges two largely separate threads in causal inference: causal *identification*, which asks whether a query can be uniquely computed from the available data, and *partial identification*, which asks how tightly the query can be bounded when unique computation is impossible. Optimal experimental design has been extensively studied for the former; we address it for the latter.

Causal identification. Shpitser and Pearl [2006] present a sound and complete algorithm (the *ID algorithm*) for determining whether the causal effect of a subset of variables, $\boldsymbol{X}$, on a target subset, $\boldsymbol{Y}$, denoted by $P(\boldsymbol{y} \mid do(\boldsymbol{x}))$, is identifiable from observational data alone, which holds if and only if a specific graphical structure known as a *hedge* is absent. Bareinboim and Pearl [2012] extend this to settings where a specific form of experimental data is also available, introducing *z-identifiability*: a query $P(\boldsymbol{y} \mid do(\boldsymbol{x}))$ is *z-identifiable* if it can be uniquely computed from the observational distribution $P(\boldsymbol{V})$ together with distributions obtained by intervening on a designated subset of variables $\boldsymbol{Z}$. Intuitively, interventions on $\boldsymbol{Z}$ can break certain confounding structures that prevent identification from observations alone. Lee et al. [2019] further extend this to *general identifiability* (g-identifiability), where the available data is an arbitrary collection of interventional distributions of the form $\{P(\boldsymbol{V} \setminus \boldsymbol{Z}_i \mid do(\boldsymbol{Z}_i))\}_{i=0}^{m}$ rather than a single family. Kivva et al. [2022] provide a sound and complete algorithm for g-identifiability, reducing it to a sequence of classical ID sub-problems.

Experimental design. Akbari et al. [2025] formulate the minimum-cost intervention design (MCID) problem for *full* identification, show that it is NP-hard, and propose an algorithm based on minimum hitting sets. Elahi et al. [2024] reformulate MCID as weighted partial MAX-SAT and integer linear programming instances, achieving orders-of-magnitude speedups in practice. Both focus on point identification. Closer to our regime, Ailer et al. [2024] sequentially design indirect experiments to narrow bound gaps in nonlinear, confounded IV settings; we instead study ex-ante subset selection in discrete graphs, with graphical certificates of zero potency.

Partial identification and causal bounds. Balke and Pearl [1997] derive sharp (i.e. tightest possible) bounds on binary causal effects using linear programming over response-type variables; their approach is restricted to discrete, in particular binary, settings with a specific graph structure. Tian and Pearl [2000] provide closed-form bounds for the probabilities of causation (PNS, PN, PS) for a single binary treatment and outcome, also requiring discrete variables. Zhang et al. [2022] bound counterfactual queries from arbitrary combinations of observational and interventional distributions via canonical SCMs. Duarte et al. [2024] further generalize to arbitrary causal and counterfactual queries in discrete settings via polynomial programming. Sachs et al. [2023] characterize conditions under which these polynomial programs reduce to linear programs. Shridharan and Iyengar [2023b] extend this class with response-type aggregation to prune the underlying LP. Shridharan and Iyengar [2023a] further show that on quasi-Markovian graphs [Zaffalon et al., 2020], the polynomial degree reduces to the number of *intervened* c-components, yielding LP or QP formulations when at most two c-components are intervened upon. Like ours, the latter exploits a c-factor decomposition [Tian and Pearl, 2002], but to reduce the structural complexity of the inner program on a restricted graph class; we instead use c-factor structure (Section 6.1) to certify experiments as uninformative ex ante on general ADMGs. In a related vein, Li et al. [2026] characterize when sharp bounds have width at most 2ϵ for binary probabilities of causation.

2 Preliminaries

Notation. Random variables are denoted by capital letters and their values by lowercase letters; bold denotes sets. For a random variable X, we write $supp(X)$ to refer to its support. When X is a binary variable, i.e. $|supp(X)| = 2$, we use x for the case $X = 1$ (or `true`) and x' for $X = 0$ (or `false`).

Structural Causal Model (SCM). An SCM is a tuple $\mathcal{M} = (\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P(\boldsymbol{U}))$ of observed (endogenous) variables $\boldsymbol{V}$, latent (exogenous) variables $\boldsymbol{U}$, structural equations $\boldsymbol{F} = \{f_V\}_{V \in \boldsymbol{V}}$, where each f_V determines the value of V as a function of its (latent and observed) parents.

Acyclic Directed Mixed Graph (ADMG). The causal structure of an SCM is summarized by an ADMG $\mathcal{G} = (\boldsymbol{V}, E^d, E^b)$ over observed variables, with directed edges $V_i \to V_j \in E^d$ indicating that V_j structurally depends on V_i, and bidirected edges $V_i \leftrightarrow V_j \in E^b$ indicating a hidden common cause of V_i and V_j. The directed part is acyclic. For a node $V \in \boldsymbol{V}$, $\boldsymbol{pa}(V)$, $\boldsymbol{ch}(V)$, and $\boldsymbol{anc}(V)$ denote its parents, children, and ancestors, respectively; these extend to sets naturally. In the figures of this paper, hidden common causes are shown as dashed nodes with directed edges to their children; this is equivalent to bidirected edges in the ADMG representation.

An *intervention* on $\boldsymbol{X} \subseteq \boldsymbol{V}$, denoted by $do(\boldsymbol{X} = \boldsymbol{x})$, replaces each structural equation f_X for $X \in \boldsymbol{X}$ with the constant value $x \in supp(X)$, removing all incoming edges to $\boldsymbol{X}$ in $\mathcal{G}$. The resulting *post-intervention distribution* on $\boldsymbol{Y} \subseteq \boldsymbol{V} \setminus \boldsymbol{X}$, $P(\boldsymbol{Y}_{\boldsymbol{x}}) = P(\boldsymbol{Y} \mid do(\boldsymbol{X} = \boldsymbol{x}))$, may differ from the observational conditional $P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ whenever hidden confounders are present. We write $P(\boldsymbol{Y}|do(\boldsymbol{X}))$ to denote the family of post-intervention distributions indexed by $\boldsymbol{x} \in supp(\boldsymbol{X})$. A *counterfactual* is a statement that involves contradictory interventions. For example, in the binary case, a counterfactual probability may take the form $P(y_x, y'_{x'})$ for variables X, Y. In this paper, we will only consider counterfactuals that are conjunctive statements like the above, where each conjunct is a counterfactual *world*. We denote the set of all counterfactual worlds appearing in a given statement by $\mathfrak{C}$. In the example above, $|\mathfrak{C}| = 2$, with interventions on $X^{(\mathfrak{c})}$ and outcomes on $Y^{(\mathfrak{c})}$ for each world $\mathfrak{c} \in \mathfrak{C}$.

Discrete variables and response-type variables. We assume all variables in $\boldsymbol{V}$ are discrete with known finite supports. Since each structural equation f_V maps finitely many parent configurations to a finite output, it can only encode a finite number of distinct input-output behaviors. For each latent disturbance U_k, these behaviors over the children of U_k are indexed by a discrete *response-type* R_k. The collection $\boldsymbol{R} = \{R_k\}_{U_k \in \boldsymbol{U}}$ is finite, and a fixed $\boldsymbol{r} \in supp(\boldsymbol{R})$ fully determines the model (including all counterfactuals). Replacing latent variables $\boldsymbol{U}$ with $\boldsymbol{R}$ yields an equivalent model with no loss of generality [Duarte et al., 2024, Proposition 1]. Given $(\mathcal{G}, P(\boldsymbol{V}), supp(\boldsymbol{V}))$ and a target query, the polynomial-programming framework of Duarte et al. [2024] constructs a program over $\boldsymbol{R}$ whose optima are sharp bounds on the query. See Appendices B (formal construction), C (primer), and A (notation table).

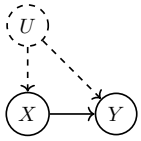
3 Causal Bounds and Epistemic Potency

Setup. We are given an ADMG $\mathcal{G}$ over observed variables $\boldsymbol{V}$ (all discrete) with distribution $P(\boldsymbol{V})$ and a causal query θ, which may correspond to a post-intervention, a counterfactual, or a functional thereof. Note that using the framework of Duarte et al. [2024], the sharp observational bounds $L_\theta \le \theta \le U_\theta$ are obtained by optimizing a polynomial program over $\boldsymbol{R}$.

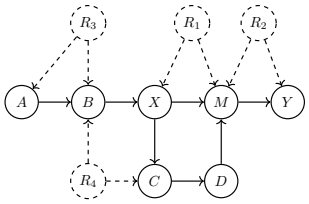
We define $\mathcal{A} := \{(y_x), (y_x, y'_{x'}), \dots\}$ as a set of *candidate experiments*, i.e., arbitrary polynomial functions of the response-type variables $\boldsymbol{R}$. We use the term broadly: $\mathcal{A}$ may contain physically realizable experiments such as randomized controlled trials and interventions [Raghavan and Bareinboim, 2025], as well as structural assumptions defensible in the application domain, such as monotonicity. The polynomial program treats both uniformly as constraints on $\boldsymbol{R}$. For any $a \subseteq \mathcal{A}$ with outcomes p_a, adding the constraints $f_a(\boldsymbol{R}) = p_a$ to the polynomial program yields new sharp bounds, denoted by $U_\theta(a, p_a)$ (upper) and $L_\theta(a, p_a)$ (lower). We write U_θ and L_θ for the bounds without any experimental constraint (i.e., $a = \emptyset$). We then define the *worst-case bound width* as follows:

$$\mathcal{W}_\theta(a) := \max_{p_a \in \prod_{\tilde{a} \in a}[l_{p_{\tilde{a}}}, u_{p_{\tilde{a}}}]} \big(U_\theta(a, p_a) - L_\theta(a, p_a)\big),$$

where $l_{p_{\tilde{a}}}$ and $u_{p_{\tilde{a}}}$ are the sharp observational bounds on $p_{\tilde{a}}$, itself a causal query. Intuitively, $\mathcal{W}_\theta(a)$ asks: if we commit to a now, how wide could the bounds still be once its outcome arrives—a worst-case width that we can promise in advance. We take the maximum because bounds must hold



(a) An ADMG with two observed variables and a hidden confounder.



(b) An ADMG with seven observed binary variables and four hidden common causes.

Figure 1: Two example ADMGs.

regardless of the experimental outcome; any smaller choice could yield a bound width that is exceeded by the realized outcome.[1] We denote the observational bound width as $\mathcal{W}_\theta := \mathcal{W}_\theta(\emptyset) = U_\theta - L_\theta$.

For illustration, Figure 1b shows an ADMG with $\boldsymbol{V} = \{A, B, X, M, Y, C, D\}$ (all binary). Restricting $\mathcal{A}$ to single-world interventions $P(\boldsymbol{W} \mid do(\boldsymbol{Z}{=}\boldsymbol{z}))$ with $\boldsymbol{W} \cap \boldsymbol{Z} = \emptyset$, each $V \in \boldsymbol{V}$ falls into exactly one of three roles: intervened with a chosen value, measured as an outcome, or neither. This results in $|\mathcal{A}| = \prod_{V \in \boldsymbol{V}} \big(|supp(V)| + 2\big) = 4^7 = 16{,}384$ candidate experiments.

Example 3.1 (Observational bounds on PNS)**.** *Consider a binary treatment X with causal effect on a binary outcome Y, confounded by a hidden variable U. The ADMG is shown in Figure 1a. Our query is the probability of necessity and sufficiency PNS $= P(y_x, y'_{x'})$, where y_x denotes $Y{=}y \mid do(X{=}x)$.*

Tian and Pearl [2000] showed that the sharp bounds for the PNS are:

$$\max\begin{Bmatrix} 0,\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), P(y_x) - P(y) \end{Bmatrix} \leq PNS \leq \min\begin{Bmatrix} P(y_x),\ P(x,y) + P(x',y'), \\ P(y'_{x'}),\ P(y_x) - P(y_{x'}) + P(x,y') + P(x',y) \end{Bmatrix}.$$

In this example, two candidate experiments are available: $\tilde{a}_1 : y_x$ and $\tilde{a}_2 : y_{x'}$. Since $P(y_x)$ and $P(y_{x'})$ are not identifiable from observational data, only the purely observational terms in the above bounds remain, giving $0 \leq PNS \leq P(x,y) + P(x',y')$. For instance, by assuming $P(x,y) = 0.2, P(x,y') = 0.1, P(x',y) = 0.3$, and $P(x',y') = 0.4$, we get $\mathcal{W}_\theta(\emptyset) = 0.6$.

Definition 3.2 (Epistemic Potency)**.** *For any $a \subseteq \mathcal{A}$ and a given causal query θ, we define $pot_\theta(a) := \mathcal{W}_\theta - \mathcal{W}_\theta(a)$. If $pot_\theta(a) = 0$, we call a* useless*. For a single experiment $\tilde{a} \in \mathcal{A}$, we write $pot_\theta(\tilde{a})$ and $\mathcal{W}_\theta(\tilde{a})$ as shorthand for $pot_\theta(\{\tilde{a}\})$ and $\mathcal{W}_\theta(\{\tilde{a}\})$, respectively.*

Potency measures how much the worst-case bound width on a causal query θ is reduced by performing experiment a. Since $\mathcal{W}_\theta(a)$ is a width we can promise before seeing any outcome, $pot_\theta(a)$ is the tightening we can promise in advance. The following result establishes that potency is always non-negative. In words, experiments can only help, never hurt.

Lemma 3.3. *For any $a \subseteq \mathcal{A}$, $pot_\theta(a) \geq 0$.*

This is because adding experimental constraints restricts the set of feasible response-type configurations, which can only tighten the resulting bounds. All detailed proofs are in Appendix H. The next result gives a useful characterization of zero potency that will underpin our pruning criteria: an experiment has zero potency if and only if some feasible outcome leaves both bounds unchanged.

Lemma 3.4. *For any $a \subseteq \mathcal{A}$, $pot_\theta(a) = 0$ if and only if $\exists\, p_a^\dagger \in \prod_{\tilde{a} \in a}[l_{p_{\tilde{a}}}, u_{p_{\tilde{a}}}]$ s.t. $U_\theta(a, p_a^\dagger) = U_\theta$ and $L_\theta(a, p_a^\dagger) = L_\theta$.*

Example 3.5. *Consider the setting of Example 3.1. (1) Performing $\tilde{a}_1$: From [Tian and Pearl, 2000], we obtain that $p_{\tilde{a}_1} \in [P(x,y),\ 1 - P(x,y')] = [0.2, 0.9]$. (2) Performing $\tilde{a}_2$: Similarly, we have $p_{\tilde{a}_2} \in [0.3, 0.6]$. For each, we obtain the following bounds:*

$$U_\theta(\tilde{a}_1, p_{\tilde{a}_1}) - L_\theta(\tilde{a}_1, p_{\tilde{a}_1}) = \begin{cases} p_{\tilde{a}_1} & p_{\tilde{a}_1} \in [0.2, 0.5], \\ 0.5 & p_{\tilde{a}_1} \in (0.5, 0.6], \\ 1.1 - p_{\tilde{a}_1} & p_{\tilde{a}_1} \in (0.6, 0.9]. \end{cases} \quad U_\theta(\tilde{a}_2, p_{\tilde{a}_2}) - L_\theta(\tilde{a}_2, p_{\tilde{a}_2}) = \begin{cases} 0.1 + p_{\tilde{a}_2} & p_{\tilde{a}_2} \in [0.3, 0.4], \\ 0.5 & p_{\tilde{a}_2} \in (0.4, 0.5], \\ 1 - p_{\tilde{a}_2} & p_{\tilde{a}_2} \in [0.5, 0.6]. \end{cases}$$

This implies the worst-case widths are $\mathcal{W}_\theta(\tilde{a}_1) = 0.5$ and $\mathcal{W}_\theta(\tilde{a}_2) = 0.5$.

(3) Performing both: Let $a_3 = \{\tilde{a}_1, \tilde{a}_2\}$. Maximizing $U_\theta(a_3, p_{\tilde{a}_1}, p_{\tilde{a}_2}) - L_\theta(a_3, p_{\tilde{a}_1}, p_{\tilde{a}_2})$ over $(p_{\tilde{a}_1}, p_{\tilde{a}_2}) \in [0.2, 0.9] \times [0.3, 0.6]$ yields $\mathcal{W}_\theta(a_3) = 0.4$. Therefore, $pot_\theta(\tilde{a}_1) = 0.1$, $pot_\theta(\tilde{a}_2) = 0.1$, $pot_\theta(a_3) = 0.2$. Thus, a_3 strictly outperforms either individual experiment, illustrating that optimal design requires reasoning over sets *of candidates. See Figure 6 in Appendix F.*

[1] Possible alternatives are briefly discussed in Appendix D.

4 The Max-Potency Problem

Here, we introduce our problem. Consider a cost function $c : 2^{\mathcal{A}} \to [0, \infty]$ that assigns a nonnegative cost to any combination of experiments in $\mathcal{A}$, along with a budget B. This function encodes the analyst's burden uniformly across both physical experiments and structural assumptions in $\mathcal{A}$—running an experiment or committing to an assumption. We allow infinite cost for impossible experiments, and assume at least one $a \subseteq \mathcal{A}$ has cost at most B. The *max-potency problem* is given by

$$\max_{a \subseteq \mathcal{A}} pot_\theta(a) \quad \text{s.t.} \quad c(a) \leq B.$$

Theorem 4.1. *The max-potency problem is NP-hard.*

We reduce 0-1 knapsack to max-potency by constructing, for each knapsack item, a disjoint confounded treatment-outcome cell whose per-cell potency can be tuned to any target value in a non-degenerate range. Disjointness makes the districts independent, and the resulting potency is additive across cells, so the max-potency optimum coincides with the knapsack optimum. See Appendix E.

Example 4.2. *Consider the ADMG in Figure 1a, with the total budget $B = 1$ and each of $\tilde{a}_1, \tilde{a}_2$ costs 1; the combined experiment $a_3 = \{\tilde{a}_1, \tilde{a}_2\}$ costs 2. The next table summarizes the resulting potencies.*

Experiment set a	$\mathcal{W}_\theta(a)$	$pot_\theta(a)$	*Cost*
$\emptyset, \tilde{a}_1, \tilde{a}_2, a_3$	$0.6, 0.5, 0.5, 0.4$	$0, 0.1, 0.1, 0.2$	$0, 1, 1, 2$

While a_3 achieves the largest potency, it is infeasible under $B = 1$. Both $\tilde{a}_1$ and $\tilde{a}_2$ are optimal, each guaranteeing a reduction of 0.1.

For completeness, Appendix G.4 gives an exact enumeration algorithm (Algorithm 5) that skips useless candidates identified by the pruning criteria of Section 6.

5 Constructing the Polynomial Program

We compute the bounds L_θ and U_θ via the polynomial-programming framework of Duarte et al. [2024]; we summarize the construction here and defer algorithmic details to Appendix G.1.

Base program. Given a canonical ADMG $\mathcal{G}$, an observational distribution $P(\boldsymbol{V})$, supports $supp(\boldsymbol{V})$, and a target query θ, Duarte et al. [2024] produce a polynomial program $\mathcal{P}_\emptyset = (\mathcal{T}, \mathcal{C}_\emptyset)$ whose decision variables are the response-type probabilities $\{P(\boldsymbol{r}) : \boldsymbol{r} \in supp(\boldsymbol{R})\}$. *Target polynomial* $\mathcal{T}$ expresses the causal query θ in the decision variables. That is, $\mathcal{T} = \sum_{\boldsymbol{r} \in h(\theta)} P(\boldsymbol{r})$, where $h(\theta) \subseteq supp(\boldsymbol{R})$ collects the configurations under which θ holds (Definition 6.1). *Constraint set* $\mathcal{C}_\emptyset$ comprises the probability axioms $P(\boldsymbol{r}) \geq 0$ and $\sum_{\boldsymbol{r}} P(\boldsymbol{r}) = 1$, together with the observational constraint, i.e., $\sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(\boldsymbol{r}) = P(\boldsymbol{v})$ for each $\boldsymbol{v} \in supp(\boldsymbol{V})$. Minimizing (maximizing) $\mathcal{T}$ subject to $\mathcal{C}_\emptyset$ yields the sharp observational lower (upper) bound on θ.

Adding experiments. For $a \subseteq \mathcal{A}$, each single experiment $\tilde{a} \in a$ with outcome $p_{\tilde{a}}$ contributes a polynomial constraint of the form: $f_{\tilde{a}}(\boldsymbol{R}) := \sum_{\boldsymbol{r} \in h(\tilde{a})} P(\boldsymbol{r}) = p_{\tilde{a}}$. That is, $f_{\tilde{a}}$ aggregates the probability mass across configurations compatible with $\tilde{a}$, which the data fixes to $p_{\tilde{a}}$. We write $p_a = (p_{\tilde{a}})_{\tilde{a} \in a}$ for the collection of outcomes. Setting $\mathcal{C}_a := \mathcal{C}_\emptyset \cup \left\{ f_{\tilde{a}}(\boldsymbol{R}) = p_{\tilde{a}} : \tilde{a} \in a \right\}$ gives the augmented program $\mathcal{P}_a = (\mathcal{T}, \mathcal{C}_a)$, whose optima are denoted by $L_\theta(a, p_a)$ and $U_\theta(a, p_a)$. The quantity $\mathcal{W}_\theta(a)$ is therefore obtained by maximizing $U_\theta(a, p_a) - L_\theta(a, p_a)$ over feasible values of p_a. See Appendix G.2 for a coupled-program formulation, which computes this worst-case width without explicitly enumerating p_a.

6 Pruning Useless Experiments

The max-potency problem is computationally challenging for three reasons: (1) finding the optimal subset is NP-hard even if pot_θ were easy to evaluate; (2) evaluating pot_θ requires solving polynomial programs; (3) $|\mathcal{A}|$ grows exponentially in $|\boldsymbol{V}|$. We cannot easily circumvent (1) or (2), but (3) offers room: we can identify $\mathcal{A}^\dagger \subseteq \mathcal{A}$ such that $\forall \tilde{a}^\dagger \in \mathcal{A}^\dagger : pot_\theta(\tilde{a}^\dagger) = 0$, a set of useless experiments, *without* solving any polynomial program.

6.1 Relevant Response-Types

To find and filter out the useless experiments, we exploit a result of Duarte et al. [2024] that shows if every directed path from a response-type to an outcome variable of $\tilde{a}$ passes through an intervened variable, that response-type is not needed for expressing $\tilde{a}$. We extend this result and show that the base program $\mathcal{P}_\emptyset$ introduced in Section 5 optimizes only over a subset of response-type variables $\boldsymbol{R}^*_\theta \subseteq \boldsymbol{R}$. Consequently, any experiment whose relevant response-types lie entirely outside $\boldsymbol{R}^*_\theta$ must have zero potency.

Strategy overview. We proceed in four steps: (1) identify *query-relevant* response-types $\boldsymbol{R}_\theta$ (those with an unblocked directed path to a query outcome); (2) expand via the *district hull* to $\boldsymbol{R}^*_\theta$ (all response-types in any district intersecting $\boldsymbol{R}_\theta$); (3) show via a Cartesian product decomposition and district factorization that $\mathcal{P}_\emptyset$ depends only on $\boldsymbol{R}^*_\theta$; (4) show that any experiment $\tilde{a}$ with $R^\star(\tilde{a}) \cap \boldsymbol{R}^*_\theta = \emptyset$ has zero potency, where $R^\star(\tilde{a})$ denotes the response-type variables relevant for $\tilde{a}$.

Definition 6.1 (Compatible Response-Types)**.** *Let $\tilde{a}$ be a joint statement of counterfactual worlds $\mathfrak{C}$, each $\mathfrak{c} \in \mathfrak{C}$ specifying interventions $\boldsymbol{X}^{(\mathfrak{c})}$ and outcomes $\boldsymbol{Y}^{(\mathfrak{c})}$. The compatible configurations are:*

$$h(\tilde{a}) := \big\{\boldsymbol{r} \in \mathit{supp}(\boldsymbol{R}) : \forall \mathfrak{c} \in \mathfrak{C},\ \forall V \in \boldsymbol{Y}^{(\mathfrak{c})},\ F_V^{(\mathfrak{c})}(\boldsymbol{r}) = v\big\},$$

where $F_V^{(\mathfrak{c})}(\boldsymbol{r})$ recursively evaluates the value of V in world $\mathfrak{c}$, using the intervention value if $V \in \boldsymbol{X}^{(\mathfrak{c})}$ and otherwise applying the structural equation.

In words, $h(\tilde{a})$ is the set of response-type configurations under which $\tilde{a}$ holds, so the probability of $\tilde{a}$ is the total mass on that set: $P(\tilde{a}) = \sum_{r \in h(\tilde{a})} P(r)$. This sum is how both queries and experiments enter the polynomial program (Section 5), but in different roles. For the query θ, it is the objective $\mathcal{T}$ that we minimize and maximize. For an experiment $\tilde{a}$, it is set equal to the experiment's outcome, which adds the constraint $\sum_{r \in h(\tilde{a})} P(r) = p_{\tilde{a}}$. Because experiments are chosen before they are run, $p_{\tilde{a}}$ is not yet known and is treated as a parameter ranging over its feasible values.

Example 6.2. *In Figure 1a, X and Y are binary. Let the tuple (R_X, R_Y) be our response-type replacing U. $R_X \in \{x', x\}$ encodes the constant value taken by X. R_Y encodes the function $\{x', x\} \to \{y', y\}$ as the pair $\big(f_Y(x'),\ f_Y(x)\big) \in \{(y', y'),\ (y', y),\ (y, y'),\ (y, y)\}$. The $\tilde{a} = \mathit{PNS} = P(y_x, y'_{x'})$ requires $f_Y(x, r_Y) = y$ and $f_Y(x', r_Y) = y'$, which uniquely selects $r_Y = (y', y)$ (the* complier*); r_X is unconstrained, thus $h(\mathit{PNS}) = \big\{(x',\ (y', y)),\ (x,\ (y', y))\big\}$.*

Definition 6.3 (Districts $\boldsymbol{D}$, District-Compatible Response-Types)**.** *The* districts $\boldsymbol{D} = \{\boldsymbol{D}_1, \ldots, \boldsymbol{D}_c\}$ *are the connected components of the bidirected graph obtained by replacing each $R \in \boldsymbol{R}$ with a bidirected edge between its children. $\boldsymbol{D}$ partitions $\boldsymbol{V} \cup \boldsymbol{R}$. For any realization $\boldsymbol{v} \in \mathit{supp}(\boldsymbol{V})$ and subset of districts $\boldsymbol{D}' \subseteq \boldsymbol{D}$ with observed variables $\boldsymbol{V}_{\boldsymbol{D}'} = \boldsymbol{D}' \cap \boldsymbol{V}$ and response-types $\boldsymbol{R}_{\boldsymbol{D}'} = \boldsymbol{D}' \cap \boldsymbol{R}$, let*

$$h_{\boldsymbol{D}'}(\boldsymbol{v}) := \big\{\boldsymbol{r}_{\boldsymbol{D}'} \in \mathit{supp}(\boldsymbol{R}_{\boldsymbol{D}'}) : \forall V_i \in \boldsymbol{V}_{\boldsymbol{D}'},\ f_{V_i}\big(\boldsymbol{pa}_{V_i}(\boldsymbol{v}),\ \boldsymbol{r}_{\boldsymbol{D}'}\big) = v_i\big\},$$

where $\boldsymbol{pa}_{V_i}(\boldsymbol{v})$ denotes the values assigned to V_i's parents by $\boldsymbol{v}$. For a single district $\boldsymbol{D}_k$, we write $h_k(\boldsymbol{v}) := h_{\boldsymbol{D}_k}(\boldsymbol{v})$.

Intuitively, a district is a maximal set of variables tied together by shared latent confounding; the observational constraints couple the response-types within a district but factorize across districts (Lemmas 6.5 and 6.7).

Remark 6.4. *In Definition 6.3, $h_{\boldsymbol{D}'}(\boldsymbol{v})$ is a special case of $h(\tilde{a})$ in Definition 6.1, where (i) the argument is $\tilde{a} = \boldsymbol{v}$, and (ii) the function is restricted to the response types in $\boldsymbol{D}'$, i.e., $\boldsymbol{R}_{\boldsymbol{D}'}$. Note that an observation can be viewed as a special type of experiment corresponding to an intervention on the empty set, $\emptyset$, yielding $\boldsymbol{V} = \boldsymbol{v}$ (in a single counterfactual world, i.e., $|\mathfrak{C}| = 1$). Because every parent value is pinned by $\boldsymbol{v}$, the recursion $F_{V'}^{(\mathfrak{c})}$ collapses to a single $f_{V'}$. Moreover, it holds that $h(\boldsymbol{v}) = h_{\boldsymbol{D}}(\boldsymbol{v})$. However, in general, $h(\tilde{a}) \neq h_{\boldsymbol{D}}(\tilde{a})$ for an arbitrary experiment $\tilde{a}$.*

Next, using the fact that f_V depends only on the response-types of V's district, we show that $h(\boldsymbol{v})$ can be decomposed into district-compatible response-types. See Appendix F.1 for an example.

Lemma 6.5 (Cartesian Product Decomposition)**.** *For any $\boldsymbol{v} \in \mathit{supp}(\boldsymbol{V})$: $h(\boldsymbol{v}) = \prod_{\boldsymbol{D}_k \in \boldsymbol{D}} h_k(\boldsymbol{v})$.*

Definition 6.6 (C-Factor [Tian and Pearl, 2002])**.** *For district $\boldsymbol{D}_k$ and a topological order $V_1 \prec \cdots \prec V_n$, we define $Q_k(\boldsymbol{v}) := Q_{\boldsymbol{D}_k}(\boldsymbol{v}) := \prod_{\{i : V_i \in \boldsymbol{D}_k \cap \boldsymbol{V}\}} P(v_i \mid \boldsymbol{v}^{(i-1)})$, where $\boldsymbol{V}^{(i-1)} := \prod_{j<i} V_j$ and $\boldsymbol{V}^{(0)} = \emptyset$. The joint distribution factorizes as $P(\boldsymbol{v}) = \prod_{k=1}^{c} Q_k(\boldsymbol{v})$.*

The next lemma connects c-factors to response-type sums, providing an algebraic bridge needed to eliminate the irrelevant districts.

Lemma 6.7. *For any $\boldsymbol{D}' \subseteq \boldsymbol{D}$ and any $\boldsymbol{v} \in supp(\boldsymbol{V})$, we can decompose $\prod_{\boldsymbol{D}_k \in \boldsymbol{D}'} Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}_{\boldsymbol{D}'} \in h_{\boldsymbol{D}'}(\boldsymbol{v})} P(\boldsymbol{r}_{\boldsymbol{D}'})$.*

Definition 6.8 (Relevant Response-Types, $\boldsymbol{R}_\theta$, District Hull, $\boldsymbol{R}^*_\theta$)**.** *$R^\star(\tilde{\mathrm{a}})$ is the set of response-types with at least one directed path to an outcome not passing through any intervened variables in any counterfactual world $\mathfrak{c} \in \mathfrak{C}$ of $\tilde{\mathrm{a}}$. We write $\boldsymbol{R}_\theta = R^\star(\theta)$ to denote the relevant response types of the query θ. Let $\boldsymbol{D}^*_\theta$ denote the set of districts with non-empty overlap with $\boldsymbol{R}_\theta$. The* district hull *is given by $D^*(\boldsymbol{R}_\theta) := \cup_{\boldsymbol{D}_k \in \boldsymbol{D}^*_\theta} \boldsymbol{D}_k \subseteq \boldsymbol{V} \cup \boldsymbol{R}$ and $\boldsymbol{R}^*_\theta = D^*(\boldsymbol{R}_\theta) \cap \boldsymbol{R}$.*

Intuitively, $\boldsymbol{R}_\theta$ contains the latent mechanisms that can still influence the query's outcomes once its interventions are in place; the hull adds every response-type that shares a district with one of them, since the observational constraints tie these together. It has been shown by [Duarte et al., 2024, Proposition 3] that the relevant response-types $R^\star(\tilde{\mathrm{a}})$ are sufficient to express the causal quantity $\tilde{\mathrm{a}}$ in the polynomial program. We can therefore formulate our objective function only in terms of $\boldsymbol{R}_\theta \subseteq \boldsymbol{R}^*_\theta$.

Example 6.9 (Query-relevant response-types in Figure 1b)**.** *Let $\theta = P(Y \mid do(M))$ in the ADMG of Figure 1b. To form $\boldsymbol{R}_\theta$, we need to find $R \in \boldsymbol{R}$ such that there is a directed path from R to Y not passing through M. $R_1 \notin \boldsymbol{R}_\theta$ since all paths $R_1 \to \cdots \to Y$ pass through M. Similarly, R_3 and R_4 are not in $\boldsymbol{R}_\theta$. In contrast, $R_2 \to Y$ does not pass through M. Hence $\boldsymbol{R}_\theta = \{R_2\}$. The districts of Figure 1b are $\boldsymbol{D}_1 = \{A, B, C, R_3, R_4\}$, $\boldsymbol{D}_2 = \{X, M, Y, R_1, R_2\}$, $\boldsymbol{D}_3 = \{D\}$. Since $R_2 \in \boldsymbol{D}_2$, then $\boldsymbol{D}^*_\theta = \{\boldsymbol{D}_2\}$. Consequently, we have $D^*(\boldsymbol{R}_\theta) = \boldsymbol{D}_2$ and $\boldsymbol{R}^*_\theta = \{R_1, R_2\}$.*

Using Lemmas 6.5 and 6.7, we can show that the observational constraints, which naively involve all of $\boldsymbol{R}$, can be reformulated to involve only $\boldsymbol{R}^*_\theta$. Specifically, the contribution of response-types outside the district hull factors out and cancels, leaving a set of constraints purely described by $\boldsymbol{R}^*_\theta$.

Lemma 6.10. *For the purpose of optimizing $\mathcal{T}$, the observational constraints $P(\boldsymbol{v}) = \sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(\boldsymbol{r})$ can equivalently be replaced by $\prod_{\boldsymbol{D}_k \in \boldsymbol{D}^*_\theta} Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}^*_\theta \in h_{\boldsymbol{D}^*_\theta}(\boldsymbol{v})} P(\boldsymbol{r}^*_\theta)$, which involves only variables in $\boldsymbol{R}^*_\theta$. The left-hand side is computable from $P(\boldsymbol{V})$.*

The base program can therefore be defined only over $\boldsymbol{R}^*_\theta$.

Example 6.11 (Reducing observational constraints via c-factors in Figure 1b)**.** *Let $\theta = P(Y \mid do(M))$ in the ADMG of Figure 1b. Under the topological order $A \prec B \prec X \prec C \prec D \prec M \prec Y$, the c-factors of the districts in Example 6.9 are $Q_1(\boldsymbol{v}) = P(a)\,P(b \mid a)\,P(c \mid a, b, x)$, $Q_2(\boldsymbol{v}) = P(x \mid a, b)\,P(m \mid a, b, x, c, d)\,P(y \mid a, b, x, c, d, m)$, and $Q_3(\boldsymbol{v}) = P(d \mid a, b, x, c)$. By Lemma 6.10, the full observational constraint $P(\boldsymbol{v}) = \sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(r_1, r_2, r_3, r_4)$ reduces to $Q_2(\boldsymbol{v}) = \sum_{(r_1, r_2) \in h_2(\boldsymbol{v})} P(r_1, r_2)$: R_3 and R_4 have been divided out, so the base program $\mathcal{P}_\emptyset$ optimizes over $|supp(R_1)| \cdot |supp(R_2)|$ decision variables rather than $\prod_{i=1}^4 |supp(R_i)|$.*

Theorem 6.12. *$\forall \tilde{\mathrm{a}} \in \mathcal{A}$ with $R^\star(\tilde{\mathrm{a}}) \cap \boldsymbol{R}^*_\theta = \emptyset$, $pot_\theta(\tilde{\mathrm{a}}) = 0$.*

Theorem 6.12 reduces zero-potency certification to checking whether an experiment's relevant response-types overlap with $\boldsymbol{R}^*_\theta$. The following corollary gives a purely graphical sufficient condition for this check.

Corollary 6.13 (Interception)**.** *Let $\tilde{\mathrm{a}} \in \mathcal{A}$ have counterfactual worlds $\mathfrak{C}$. If for every $\mathfrak{c} \in \mathfrak{C}$, every directed path from $\boldsymbol{R}^*_\theta$ to $\boldsymbol{Y}^{(\mathfrak{c})}$ passes through $\boldsymbol{X}^{(\mathfrak{c})}$, then $pot_\theta(\tilde{\mathrm{a}}) = 0$.*

Example 6.14 (Zero potency)**.** *Consider experiment $P(D \mid do(C))$ in Example 6.9. The only directed path from $\boldsymbol{R}^*_\theta$ to D is $R_1 \to C \to D$, which passes through C. Hence $pot_\theta(P(D \mid do(C))) = 0$. On the other hand, the experiment $P(D \mid do(B))$ has potentially non-zero potency, since the path $R_1 \to X \to C \to D$ does not pass through B.*

Algorithm 4 in Appendix G.3 determines whether a given single-world experiment is useless by verifying the condition of Corollary 6.13. If this algorithm returns `True`, the intervention is deemed useless; otherwise, the result is inconclusive. Its computational complexity is linear in the size of the graph. For a worked example on Figure 1b, see Appendix F.2.

6.2 Identifiable Experiments

When an intervention is identifiable, it cannot provide any additional information beyond what is already contained in the observational distribution. In other words, it is useless.

Theorem 6.15. *$\forall\, \tilde{a} \in \mathcal{A}$ where $P(\tilde{a})$ is identifiable (*ID*) in $\mathcal{G}$, $pot_\theta(\tilde{a}) = 0$.*

The ID algorithm in Shpitser and Pearl [2006] can be applied to verify identifiability for single-world interventions $P(\boldsymbol{w} \mid do(\boldsymbol{z}))$. Note that cross-world counterfactuals are not ID by construction, so ID is applied only to interventions. Furthermore, the results of Theorems 6.12 and 6.15 can be applied independently, in any order, to filter out useless experiments, and Algorithm 1 exploits this fact.

Corollary 6.16 (ID-inertness)**.** *$\forall \tilde{a} \in \mathcal{A}$ where $\tilde{a}$ is* ID*: $pot_\theta(\{\tilde{a}\} \cup b) = pot_\theta(b)$ for any $b \subseteq \mathcal{A}$.*

In words, ID-pruned experiments can be removed from the subset search entirely without affecting the optimum (Appendix G.4).

Algorithm 1: GetUselessExperiments

Require: Graph $\mathcal{G}$, Query θ, Experiment set $\mathcal{A}$
Ensure: $\mathcal{A}^\dagger \subseteq \mathcal{A}$ with $\forall\, \tilde{a} \in \mathcal{A}^\dagger : pot_\theta(\tilde{a}) = 0$
  $\boldsymbol{R}^*_\theta \leftarrow D^*\big(R^\star(\theta)\big) \cap \boldsymbol{R}$; $\mathcal{A}^\dagger \leftarrow \emptyset$
  for $\tilde{a} \in \mathcal{A}$ with worlds $\mathfrak{C}_{\tilde{a}}$ **do**
    if $\forall \mathfrak{c} \in \mathfrak{C}_{\tilde{a}}$: ALG. 4$(\mathcal{G}, \boldsymbol{R}^*_\theta, \boldsymbol{Y}^{(\mathfrak{c})}, \boldsymbol{X}^{(\mathfrak{c})})$ **then**
      $\mathcal{A}^\dagger \leftarrow \mathcal{A}^\dagger \cup \{\tilde{a}\}$; **continue**
    end if
    if $|\mathfrak{C}_{\tilde{a}}| = 1$ **and** ID$(\mathcal{G}, \boldsymbol{Y}^{(\mathfrak{c})}, \boldsymbol{X}^{(\mathfrak{c})})$ **then**
      $\mathcal{A}^\dagger \leftarrow \mathcal{A}^\dagger \cup \{\tilde{a}\}$
    end if
  end for

Example 6.17 (Running Algorithm 1 on Figure 1b)**.** *Let $\theta = P(y \mid do(m))$ as in Example 6.9, so $\boldsymbol{R}^*_\theta = \{R_1, R_2\}$, and consider the candidates $\tilde{a}_1 : P(C, D \mid do(x))$, $\tilde{a}_2 : P(Y, M, X \mid do(a, b))$, $\tilde{a}_3 : P(Y, M \mid do(x))$, $\tilde{a}_4 : P(Y, M \mid do(x'))$, $\tilde{a}_5 : P(B \mid do(a))$, and $\tilde{a}_6 : P(d_{c'}, d'_c)$, i.e., testing for monotonicity along $C \to D$. Interception (Algorithm 4) certifies $\tilde{a}_1$, $\tilde{a}_5$ and $\tilde{a}_6$ as useless: every directed path from $\{R_1, R_2\}$ to their outcomes passes through the intervened variable (X, A and C, respectively). Moreover, $P(\tilde{a}_2)$ is identifiable, so $\tilde{a}_2$ is useless by Theorem 6.15 and may be dropped from any subset search (Corollary 6.16). Algorithm 1 thus returns $\mathcal{A}^\dagger = \{\tilde{a}_1, \tilde{a}_2, \tilde{a}_5, \tilde{a}_6\}$, and only $\tilde{a}_3$ and $\tilde{a}_4$ need to be evaluated numerically.*

7 Evaluation

Synthetic evaluation. We evaluate Algorithm 1 on two types of graphs: randomly generated graphs and several real-world networks selected from the *bnlearn*[2] repository. For each graph, we randomly choose a query as well as a set of $|\mathcal{A}| = 200$ candidate experiments. We report the mean fraction of candidates pruned by the two core mechanisms of Algorithm 1: the ID check, and the interception rule (Algorithm 4); throughout, $\pm$ denotes one standard deviation across simulations. For details on graph, query, and experiment generation as well as per-graph results we refer to Appendix K.

Erdős–Rényi random graphs: We sample ADMGs with $|\boldsymbol{V}| \in \{10, 15, 20, 30\}$ observed nodes and vary the confounding density $|\boldsymbol{U}|/|\boldsymbol{V}|$ from roughly 0.25 to 1.0. Across $1{,}596$ simulations, Algorithm 1 prunes on average $74.3\% \pm 19.4\%$ of candidate experiments ($33.0\% \pm 23.2\%$ via the ID check, $41.3\% \pm 38.5\%$ via interception), reaching as much as 90.6% in some simulations. Figure 2 reveals a clear pattern: as $|\boldsymbol{U}|/|\boldsymbol{V}|$ grows, interception weakens (from $70.9\% \pm 31.0\%$ at the sparsest setting to $16.5\% \pm 30.0\%$ at the densest). The fraction pruned by the ID check, in turn, increases with confounding density, since the ID algorithm is only executed on experiments that were not already pruned by interception.

bnlearn: We evaluate on 11 benchmark networks ranging from $|\boldsymbol{V}|$=8 (`ASIA`) to $|\boldsymbol{V}|$=76 (`WIN95PTS`). Across $1{,}100$ simulations, Algorithm 1 prunes on average $63.4\% \pm 12.4\%$ of candidate experiments ($47.1\% \pm 14.5\%$ via the ID check, $16.2\% \pm 17.5\%$ via interception). Figure 3 displays the pruning rates per graph. The contribution of the interception rule varies dramatically across graphs, ranging from $2.3\% \pm 4.3\%$ on `MILDEW` to $29.2\% \pm 25.5\%$ on `WIN95PTS`.

The high variability in the interception rate (e.g. $\pm 25.5\%$ on `WIN95PTS`) reflects its strong dependence on the specific topology. Figure 9 in Appendix K.5 shows the main structural driver of interception pruning. When the query is localized (small $|\boldsymbol{R}^*_\theta|/|\boldsymbol{V}|$), most experiments fail to interact with $\boldsymbol{R}^*_\theta$ and are pruned immediately; as the hull grows, this mechanism weakens monotonically. This matches the Erdős–Rényi findings: denser confounding produces larger districts, which enlarge $\boldsymbol{R}^*_\theta$ and shift

[2] `https://www.bnlearn.com/bnrepository/`

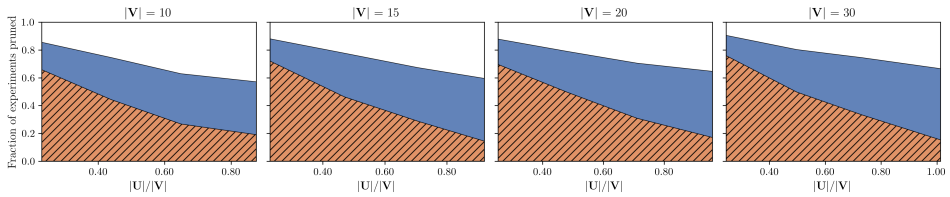


Figure 2: Erdős–Rényi graphs. For different sizes of $\boldsymbol{V}$, we plot the fraction of pruned experiments against the density $\frac{|\boldsymbol{U}|}{|\boldsymbol{V}|}$. The hatched orange area represents experiments pruned by the interception mechanism (Algorithm 4); the blue area represents those pruned by the ID check. In total, $1{,}596$ simulations were performed.

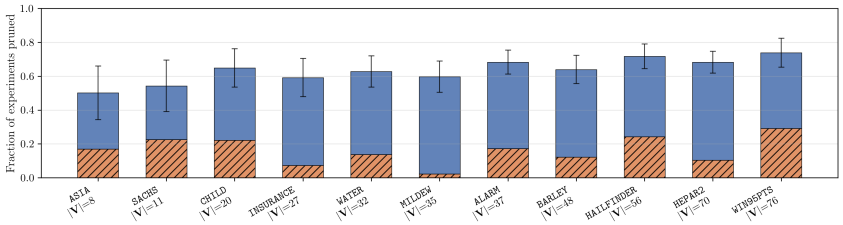


Figure 3: *bnlearn*. For each graph, we plot the mean fraction of pruned experiments. The hatched orange area represents experiments pruned by the interception mechanism (Algorithm 4); the blue area represents those pruned by the ID check. Error bars indicate ± 1 standard deviation of the total pruning rate across 100 simulations per graph. In total, $1{,}100$ simulations were performed.

the pruning burden onto the ID check. Together, the two criteria yield substantial pruning across singleton candidates without solving a single polynomial program.

Subset search space reduction: The pruning rates reported above determine the search space reduction when selecting a single optimal experiment, a practically important setting. For the general subset search over $2^{|\mathcal{A}|}$ subsets, Corollary 6.16 reduces the exponent by $|\mathcal{A}^{\dagger}_{\mathrm{ID}}|$ (Appendix G.4). Running the ID check on all candidates independently (without prior interception filtering), it certifies $61.4\% \pm 11.3\%$ on Erdős–Rényi and $60.0\% \pm 10.8\%$ on *bnlearn*—for $|\mathcal{A}| = 200$, the exponent drops from 200 to roughly 80 on average.

Runtime: Pruning inspects only the graph and the query and runs in milliseconds; what it saves is solver time. On 26 random graphs with $|\boldsymbol{V}| = 5$ (small enough for the solver to finish) and all 152 single-variable interventions they support, pruning took about 6 ms per graph, while evaluating every candidate took at least 7.4 hours of solver time against 1.1 hours for the 48 survivors (Appendix J.2), a speedup of at least $6.95\times$ that exceeds the drop in candidate count because the discarded candidates are also the slowest to solve (median 612 s versus 2.5 s).

Physical Activity and Diabetes. We demonstrate the full pipeline, pruning followed by potency computation, on observational data from NHANES 2017–2018.[3]

Graph and query: Five binary variables describe the setting: balance problems (A), falls (B), health insurance (Z), physical activity (X), and not diabetes (Y).[4] Three latent confounders could represent inner-ear disorders (U_3, confounding $A \leftrightarrow B$), healthcare access (U_1, confounding $Z \leftrightarrow X$), and genetic predisposition (U_2, confounding $X \leftrightarrow Y$). The assumed causal structure is shown in Figure 4; Appendix L reports experiment types, computational details, and robustness analyses. The query of interest is $\theta = \mathrm{ATE} = P(y_x) - P(y_{x'})$, the average treatment effect of physical activity on not getting diabetes. The observational data yields a bound width of $\mathcal{W}_\theta = 0.864$ with $\theta \in [-0.445, 0.419]$.

Experiments: In practice, an analyst need not commit the entire research budget at once; a natural workflow is to select a small optimal batch, execute it, and re-solve with the realized outcome incorporated. We consider the candidate experiments and combinations in Table 1. Algorithm 1 certifies $\tilde{a}_1$ and $\tilde{a}_2$ as useless.

[3] `https://wwwn.cdc.gov/nchs/nhanes/`

[4] We invert Y for better interpretability.

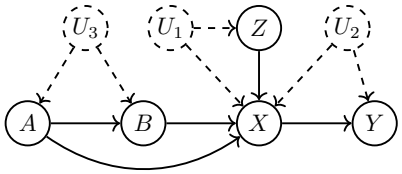


Figure 4: Assumed causal ADMG for the NHANES experiment ($n = 5{,}833$ complete cases). Solid nodes are observed; dashed nodes are latent confounders.

Table 1: Pruning and potency results. Left: cost-1 singleton experiments ($\tilde{a}_1$, $\tilde{a}_2$ pruned by ID and Interception, respectively). Right: cost-2 singletons and combinations.

	Experiment	Pruned?	pot_θ	cost
$\tilde{a}_1$	$P(X \mid do(B))$	**ID**	0	1
$\tilde{a}_2$	$P(B \mid do(A))$	**Int.**	0	1
$\tilde{a}_3$	$P(X, Y \mid do(z))$	–	0	1
$\tilde{a}_4$	$P(y_{x'}, y'_x)$	–	0.44	1

	Experiment	pot_θ	cost
$\tilde{a}_5$	$P(Y \mid do(x'))$	0.36	2
$\tilde{a}_6$	$P(Y \mid do(x))$	0.50	2
	$\{\tilde{a}_4, \tilde{a}_6\}$	0.50	3
	$\{\tilde{a}_4, \tilde{a}_5\}$	0.78	3

After computing the potencies of the remaining experiments, we can see that $\tilde{a}_3$ is useless, but was not pruned by either criterion. The solution to the max-potency problem depends on the budget B committed per round. For $B = 1$, the optimal choice is $\{\tilde{a}_4\}$. For $B = 2$, we should carry out the positive intervention $\{\tilde{a}_6\}$. For $B = 3$, the most potent combination is $\{\tilde{a}_4, \tilde{a}_5\}$—even though the positive intervention ($\tilde{a}_6$) is more potent by itself, in combination with $\tilde{a}_4$, the negative intervention ($\tilde{a}_5$) is superior. Committing to $B = 1$ and finding $P(y_{x'}, y'_x) = 0$ (intuitively, physical activity cannot cause diabetes) would yield $\theta \in [0.074, 0.419]$, after which the analyst re-solves against the updated program.

8 Conclusion

We introduced the max-potency problem—selecting cost-constrained experiments to maximally tighten bounds on a partially identifiable causal query—defined *epistemic potency* as a worst-case measure of informativeness, and showed the problem is NP-hard. Two graphical pruning criteria, a path-interception rule and an ID check, together prune 50–88% of the super-exponential candidate space for singleton selection. For subset search, ID-pruned experiments are combinatorially inert (Corollary 6.16) and can be removed entirely, reducing the exponent by roughly 60% on average. A natural extension is the sequential setting, where realized outcomes are incorporated into the base program after each round and the next optimal batch is selected against the updated bounds, as illustrated in Section 7.

Limitations. i) Our framework inherits limitations from Duarte et al. [2024]: known ADMG, discrete variables with finite support, exact $P(\boldsymbol{V})$, the computational cost of the bounding programs (Appendix J), and possibly non-sharp bounds if the nonconvex polynomial program is terminated early. ii) Our pruning criteria are sound but not complete: experiments not certified useless may still have zero potency, and interception-pruned experiments may contribute non-zero potency in subsets. iii) Our worst-case bound-width definition (Section 3) guarantees tightening regardless of outcome; under alternative criteria (Appendix D) the optimal set may differ.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and area chairs for their constructive feedback. This work received no third-party funding. The authors declare no competing interests.

References

E. Ailer, N. Dern, J. Hartford, and N. Kilbertus. Targeted sequential indirect experiment design. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

S. Akbari, J. Etesami, and N. Kiyavash. Optimal experiment design for causal effect identification. *Journal of Machine Learning Research*, 26(28):1–56, 2025. URL `http://jmlr.org/papers/v26/22-1516.html`.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

A. Balke and J. Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120. AUAI Press, 2012.

G. Duarte, N. Finkelstein, D. Knox, J. Mummolo, and I. Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793, 2024. doi: 10.1080/01621459.2023.2216909. URL https://doi.org/10.1080/01621459.2023.2216909.

S. Elahi, S. Akbari, J. Etesami, N. Kiyavash, and P. Thiran. Fast proxy experiment design for causal effect identification. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 51254–51283. Curran Associates, Inc., 2024. doi: 10.52202/079017-1622. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5bd9fbb3a5a985f80c16ddd0ec1dfc43-Paper-Conference.pdf.

C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. doi: 10.1111/j.0006-341X.2002.00021.x.

S. Greenland and J. M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419, 09 1986. ISSN 0300-5771. doi: 10.1093/ije/15.3.413. URL https://doi.org/10.1093/ije/15.3.413.

Y. Kivva, E. Mokhtarian, J. Etesami, and N. Kiyavash. Revisiting the general identifiability problem. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180, pages 1022–1030. PMLR, 2022.

S. Lee, J. D. Correa, and E. Bareinboim. General identifiability with arbitrary surrogate experiments. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115, pages 389–398. PMLR, 2019.

A. Li, S. Mueller, X. Shu, and J. Pearl. ϵ-Identifiability of Causal Quantities. In E. Khan, Y. Li, A. Solin, and A. Ramdas, editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 4978–4986. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/li26j.html.

S. Mueller and J. Pearl. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023. doi: 10.1515/jci-2022-0050. URL https://doi.org/10.1515/jci-2022-0050.

J. Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121:93–149, 1999.

A. Raghavan and E. Bareinboim. Counterfactual realizability. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. Spotlight.

M. C. Sachs, G. Jonzon, A. Sjölander, and E. E. Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2):567–576, 2023. doi: 10.1080/10618600.2022.2071905. URL https://doi.org/10.1080/10618600.2022.2071905.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1219–1226. AAAI Press, 2006.

M. Shridharan and G. Iyengar. Causal bounds in quasi-Markovian graphs. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31675–31692. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/shridharan23a.html.

M. Shridharan and G. Iyengar. Scalable computation of causal bounds. *Journal of Machine Learning Research*, 24(237):1–35, 2023b.

J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):287–313, Oct. 2000. ISSN 1573-7470. doi: 10.1023/A:1018912507879. URL https://doi.org/10.1023/A:1018912507879.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, page 567–573, USA, 2002. American Association for Artificial Intelligence. ISBN 0262511290.

M. Zaffalon, A. Antonucci, and R. Cabañas. Structural causal models are (solvable by) credal networks. In M. Jaeger and T. D. Nielsen, editors, *Proceedings of the 10th International Conference on Probabilistic Graphical Models*, volume 138 of *Proceedings of Machine Learning Research*, pages 581–592. PMLR, 23–25 Sep 2020. URL https://proceedings.mlr.press/v138/zaffalon20a.html.

J. Zhang, J. Tian, and E. Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26548–26558. PMLR, 2022.

Appendix

A Notation

Table 2 summarizes the notation used throughout the paper. The Level column separates element-level from set-level symbols: a single candidate experiment is written $\tilde{a}$, a set of experiments a, and the candidate pool $\mathcal{A}$; for singletons we use the shorthand of Definition 3.2.

Table 2: Summary of notation.

Symbol	Level	Meaning
X; x	—	random variable; realized value (x/x' for binary true/false)
$\boldsymbol{X}$, $\boldsymbol{x}$	set	bold symbols denote sets of variables/values
$supp(X)$	—	support of X
$\mathcal{M} = (\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{F}, P(\boldsymbol{U}))$	—	structural causal model
$\mathcal{G} = (\boldsymbol{V}, E^d, E^b)$	—	ADMG: directed (E^d) and bidirected (E^b) edges
$\boldsymbol{pa}(V)$, $\boldsymbol{ch}(V)$, $\boldsymbol{anc}(V)$	—	parents, children, ancestors of V
$do(\boldsymbol{X} = \boldsymbol{x})$; $P(\boldsymbol{Y}_{\boldsymbol{x}})$	—	intervention; post-intervention distribution
$\mathfrak{C}$; $\mathfrak{c}$	—	counterfactual worlds of a statement; a single world
$\boldsymbol{X}^{(\mathfrak{c})}$, $\boldsymbol{Y}^{(\mathfrak{c})}$	set	intervened variables and outcomes of world $\mathfrak{c}$
R_k; $\boldsymbol{R}$; $\boldsymbol{r}$	—	response type of disturbance U_k; all response types; a configuration
θ	—	target causal query
$h(\cdot)$	—	response-type configurations compatible with a statement
$R^\star(\cdot)$	—	relevant response types of a causal quantity
$\boldsymbol{R}_\theta$; $\boldsymbol{D}^*_\theta$; $\boldsymbol{R}^*_\theta$	set	query-relevant response types; their districts; district hull
Q_k	—	c-factor of district $\boldsymbol{D}_k$
$\mathcal{A}$	pool	set of all candidate experiments
$\tilde{a} \in \mathcal{A}$	element	a single candidate experiment
$a \subseteq \mathcal{A}$	set	a set of experiments
$\mathcal{A}^\dagger \subseteq \mathcal{A}$	set	experiments certified useless by pruning
$p_{\tilde{a}}$; p_a	element; set	outcome of one experiment; collection of outcomes
$[l_{p_{\tilde{a}}}, u_{p_{\tilde{a}}}]$	element	feasible range of an outcome
$f_{\tilde{a}}$	element	constraint function of an experiment
U_θ, L_θ	—	observational upper/lower bound on θ
$U_\theta(a, p_a)$, $L_\theta(a, p_a)$	set	bounds under experimental constraints
$\mathcal{W}_\theta(a)$; $\mathcal{W}_\theta$	set	worst-case width; observational width
$pot_\theta(a)$	set	epistemic potency (Definition 3.2)
$c(\cdot)$; B	set	cost function on experiment sets; budget
$\mathcal{P}_\emptyset$; $\mathcal{P}_a$	—	base and augmented polynomial programs

B Definition of Response-Type Variables

Following Duarte et al. [2024], we define response-type variables per disturbance and assume the given ADMG $\mathcal{G}$ to be *canonical* without loss of generality (see their paper for details on canonicalization).

Definition B.1 (Response-Type Variable)**.** *Let $\mathcal{G}$ be a canonical ADMG with discrete observed variables $\boldsymbol{V}$ and latent disturbances $\boldsymbol{U}$. For a disturbance $U_k \in \boldsymbol{U}$ with observed children $\boldsymbol{ch}(U_k) \subseteq \boldsymbol{V}$, declare two values $u_k, u'_k \in supp(U_k)$ equivalent, written $u_k \sim_k u'_k$, if for every $V \in \boldsymbol{ch}(U_k)$ the partially applied structural equation*

$$f_V^{(u_k)} : \prod_{P \in \boldsymbol{Pa}(V) \setminus \{U_k\}} supp(P) \;\to\; supp(V)$$

is identical under u_k and u'_k. The equivalence classes $supp(U_k)/\sim_k$ are the generalized principal strata *of U_k [Frangakis and Rubin, 2002, Balke and Pearl, 1997], and the* response-type variable R_k *associated with U_k is the discrete categorical variable whose support is exactly these classes. The induced distribution $P(R_k)$ aggregates $P(U_k)$ over each class. We write $\boldsymbol{R} = \{R_k\}_{U_k \in \boldsymbol{U}}$ for the collection of all response-type variables.*

Proposition B.2 (Duarte et al., 2024, Proposition 1)**.** *Replacing each U_k in the canonical SCM with its response-type variable R_k yields an equivalent SCM with respect to the full data law (every factual, interventional, and counterfactual joint distribution is preserved).*

C A Primer on the Polynomial-Programming Framework

This appendix gives a jargon-free walkthrough of the bounding framework of Duarte et al. [2024] on which our method builds, using the running example of Section 3; Appendix B gives the formal construction.

The idea. When a causal quantity cannot be determined exactly from the available data, we instead ask: what is the smallest and largest value it could take while remaining consistent with the data? Consider the PNS in Figure 1a, where the binary variables X and Y share an unobserved common cause U. Because X and Y are binary, U can influence them in only finitely many deterministic ways. It can fix X to either 0 or 1, and it can select one of four possible relationships between X and Y: $Y = 1$ regardless of X; $Y = 1$ exactly when $X = 1$ (the "complier" type); $Y = 1$ exactly when $X = 0$; and $Y = 0$ regardless of X. Thus U can be replaced, without loss of generality, by a discrete response type with $2 \times 4 = 8$ possible values (2 possible behaviors for X, 4 for Y). The population is then described by eight unknown probabilities, one for each type, and every observational, interventional, or counterfactual quantity can be written in terms of these probabilities.

Why this gives bounds. The PNS is the proportion of the population belonging to the complier type, summed over the two possible response types for X. The observed distribution $P(X, Y)$ restricts where this complier probability can be placed: a complier has $Y = X$, so compliers may contribute to the observed cells $(X, Y) = (1, 1)$ and $(0, 0)$, but not to $(1, 0)$ or $(0, 1)$. Therefore the total complier probability cannot exceed $P(X{=}1, Y{=}1) + P(X{=}0, Y{=}0)$, which is exactly the observational upper bound in Example 3.1. Bounding is therefore an optimization problem: maximize or minimize the causal quantity of interest subject to the constraints imposed by the observed data.

What Duarte et al. [2024] automate. Their method mechanizes these steps for general discrete causal graphs and queries: it constructs the possible response types from the graph, expresses the target query and the observed probabilities in terms of the response-type probabilities, and sends the resulting minimization or maximization problem to a global optimization solver such as Couenne. When the graph contains several independent unobserved variables, the probability of a joint response-type configuration is a *product* of their individual probabilities; the resulting optimization problem therefore contains polynomial, rather than only linear, expressions and requires a global solver.

Our work builds on this framework: an experiment contributes additional constraints of the same general form, as described in Section 5, and potency measures how much these new constraints narrow the resulting bounds.

D Alternative Criteria for $\mathcal{W}_\theta$

In Section 3, we defined $\mathcal{W}_\theta(a)$ using the worst-case (maximum) over feasible experimental outcomes. Here, we briefly discuss two alternative criteria and compare them on our running example.

Option 1: Expected width. One could assume a distribution over the unknown experimental outcome, e.g. $p_{\tilde{a}} \sim \mathrm{Uni}(l_{p_{\tilde{a}}}, u_{p_{\tilde{a}}})$, and define

$$\mathcal{W}_\theta(\tilde{a}) = \mathbb{E}[U_\theta(\tilde{a}, p_{\tilde{a}}) - L_\theta(\tilde{a}, p_{\tilde{a}})] = \frac{1}{u_{p_{\tilde{a}}} - l_{p_{\tilde{a}}}} \int_{l_{p_{\tilde{a}}}}^{u_{p_{\tilde{a}}}} U_\theta(\tilde{a}, p_{\tilde{a}}) - L_\theta(\tilde{a}, p_{\tilde{a}})\, dp$$

for a singleton $\tilde{a} \in \mathcal{A}$ as the expected width.

Option 2: Best-case width. One may change the maximum to a minimum in the original definition and define

$$\mathcal{W}_\theta(a) = \min_{p_a \in \prod_{\tilde{a} \in a} [l_{p_{\tilde{a}}}, u_{p_{\tilde{a}}}]} \big(U_\theta(a, p_a) - L_\theta(a, p_a)\big)$$

as the smallest possible bound width an experiment could yield.

Table 3: Comparison of alternative width criteria on the running example.

Criterion	$\mathcal{W}_\theta$	$\mathcal{W}_\theta(\tilde{a}_1)$	$\mathcal{W}_\theta(\tilde{a}_2)$	$\mathcal{W}_\theta(a_3)$
Expected width	0.6	0.371	0.467	0.238
Best-case width	0.6	0.2	0.4	0.0
Worst-case width	0.6	0.5	0.5	0.4

Comparison on the running example. Table 3 displays alternative definitions of $\mathcal{W}_\theta$ under the setting of Example 3.1. Notably, at $(p_{\tilde{a}_1}, p_{\tilde{a}_2}) = (0.2, 0.6)$, the bounds collapse to PNS $= 0$; the PNS is not identifiable from observational data alone, but becomes point-identified for this specific experiment realization.

Under all three criteria, a_3 yields the smallest bound width; among singleton experiments, both $\tilde{a}_1$ and $\tilde{a}_2$ are equivalent under the worst-case criterion, while $\tilde{a}_1$ is strictly preferred under the alternatives. Only the worst-case criterion provides a *guaranteed* reduction in bound width regardless of the experimental outcome, which is why we adopt it throughout this paper.

D.1 Criterion Sensitivity on the NHANES Candidates

To assess how much the choice of criterion matters in practice, we recomputed the unpruned NHANES candidates of Table 1 under all three criteria. Table 4 reports the resulting potencies.[5]

Table 4: NHANES candidates under the worst-case, expected, and best-case criteria (potency values; observational width $\mathcal{W}_\theta = 0.86$).

Candidate	Worst	Expected	Best	Cost
$\tilde{a}_3 = P(X, Y \mid do(z))$	0	0.49	0.72	1
$\tilde{a}_4 = P(y_{x'}, y'_x)$	0.44	0.44	0.52	1
$\tilde{a}_5 = P(Y \mid do(x'))$	0.36	0.36	0.36	2
$\tilde{a}_6 = P(Y \mid do(x))$	0.50	0.50	0.50	2
$\{\tilde{a}_4, \tilde{a}_6\}$	0.50	0.51	0.55	3
$\{\tilde{a}_4, \tilde{a}_5\}$	0.78	0.78	0.83	3

The criterion changes the *decision*, not merely the numbers: at $B = 1$, the worst-case criterion selects $\tilde{a}_4$, whereas the expected and best-case criteria both select $\tilde{a}_3$—an experiment whose worst-case potency is zero, so it cannot be guaranteed to narrow the bounds at all, yet which reaches 0.49 on average over its outcomes and 0.72 at its most favorable one. At $B = 3$, $\{\tilde{a}_4, \tilde{a}_5\}$ is optimal under all three criteria among the sets evaluated.

That said, the criterion is a modular choice within our framework. (i) The pruning results hold under all three criteria: our proofs show something stronger than zero worst-case potency—for pruned experiments, the bounds remain unchanged under *every* feasible outcome (Theorems 6.12 and 6.15). Pruning is therefore unaffected by the criterion; only the ranking among the survivors depends on it. (ii) For potency evaluation, the worst-case criterion is what allows collapsing the whole computation into a single polynomial program (Appendix G.2), and the collapse is specific to it. The worst case maximizes a difference between a maximum and a minimum, so all three optimizations merge into one; the best case replaces the outer maximum by a minimum and is a min–max, while the expected width is an integral over outcomes rather than an optimization at all. Under either alternative one must sweep the feasible outcomes and solve a separate program at each: the Expected and Best columns of Table 4 take hundreds to thousands of programs per candidate, against one for the Worst column.

[5]The Worst column is exact: it is the single coupled program of Appendix L.2. The Expected and Best columns admit no such collapse (see (ii) below), so they are obtained by sweeping the feasible outcomes—800 points per scalar outcome dimension, 30×30 for outcome pairs, and for $\tilde{a}_3$ a sweep of 30 sampled feasible $do(Z{=}0)$ distributions against the feasible distributions on a $1/30$ enumeration grid for $do(Z{=}1)$—and are therefore subject to discretization error.

D.2 Beyond a Single Budget: Money and Time

The budget model is equally modular. Practitioners may care about resources beyond money, such as the time needed to conduct an experiment. Concretely, let each experiment set carry a pair of costs $(c(a), t(a))$, money and time, with two budgets B_c and B_t. The max-potency problem becomes $\max_{a \subseteq \mathcal{A}} pot_\theta(a)$ subject to $c(a) \leq B_c$ *and* $t(a) \leq B_t$. Costs never enter the potency computation itself; they only decide which sets are affordable, so the only change to the solution method is that the single budget check per candidate set becomes double. The pruning procedure is likewise unaffected: it uses the graph, the query, and the candidate list, never the costs—a zero-potency experiment is useless no matter how cheap or fast it is. Since the original problem is the special case $B_t = \infty$, the extension remains NP-hard, and the computational bottleneck is unchanged. Note that our cost functions are defined over *sets* of experiments, so time need not be additive—experiments performed in parallel cost the maximum of their durations rather than the sum, and the formulation captures this as-is. In the sequential setting, which we leave as future work, time becomes even more relevant: for example, one may seek to reach a target bound width within a given deadline at least cost.

E NP-hardness of the Max-Potency Problem

In this appendix we prove Theorem 4.1 (max-potency is NP-hard) via a polynomial-time reduction from the 0-1 knapsack problem.

At a high level, the reduction maps every knapsack object to an experiment: the object's weight becomes the experiment's cost, and the object's value becomes the experiment's potency. The gadget that makes this work is the *TP-cell* defined below: a minimal ADMG with a treatment X_i, an outcome Y_i, and confounding between them, in which the query of interest is the PNS_i with its symbolic bounds from Tian and Pearl [2000]. Given a knapsack instance with values v_i, weights w_i, and budget B, we build one TP-cell per item, choose each cell's observational distribution so that the potency of the cell's single experiment $\tilde{a}_i = P(y_{i,x_i})$ encodes the item's value (Lemma E.3; the formal proof first rescales the values into the achievable range, which does not change the optimal knapsack solution), take the query to be the sum of all PNS_i, and assign $\tilde{a}_i$ cost w_i. Because distinct cells share no variables, performing several experiments tightens each cell's bounds independently, so the potency of a set of experiments is simply the sum of the per-cell potencies (Lemma E.2)—just like the value of a set of packed objects. Solving the constructed max-potency instance therefore solves the knapsack instance, and each step of the construction runs in polynomial time.

E.1 The atomic unit: TP-cells

Definition E.1 (TP-cell)**.** *A* TP-cell $\mathcal{C}_i$ *is the ADMG with binary observed variables* X_i, Y_i*, a directed edge* $X_i \to Y_i$*, and a bidirected edge* $X_i \leftrightarrow Y_i$*, equipped with an observational distribution* $P_i(X_i, Y_i)$ *parameterized by*

$$\alpha_i = P_i(x_i, y_i), \quad \beta_i = P_i(x_i, y_i'), \quad \gamma_i = P_i(x_i', y_i), \quad \delta_i = P_i(x_i', y_i'),$$

with $(\alpha_i, \beta_i, \gamma_i, \delta_i) \in \Delta$*, where*

$$\Delta := \left\{(\alpha, \beta, \gamma, \delta) \in \mathbb{R}^4_{\geq 0} : \alpha + \beta + \gamma + \delta = 1\right\}$$

denotes the probability simplex over the four joint outcomes of (X_i, Y_i)*.*

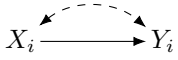


Figure 5: A TP-cell $\mathcal{C}_i$: a confounded treatment-outcome pair.

The query of interest within a single cell is the probability of necessity and sufficiency,

$$\mathrm{PNS}_i = P(y_{i,x_i}, y'_{i,x'_i}).$$

For a TP-cell, Tian and Pearl [2000] give the sharp observational bounds

$$0 \leq \mathrm{PNS}_i \leq \alpha_i + \delta_i,$$

so that the observational worst-case width is

$$\mathcal{W}_i(\emptyset) = \alpha_i + \delta_i.$$

We write $pot_i(\tilde{a}_i) = \mathcal{W}_i(\emptyset) - \mathcal{W}_i(\tilde{a}_i)$ for the per-cell potency, in line with Definition 3.2. Furthermore, we use $W_{(\cdot)}(a, p_a) := U_{(\cdot)}(a, p_a) - L_{(\cdot)}(a, p_a)$ to denote the instantiated bound width.

E.2 The construction

Given a 0-1 knapsack instance with values $v_1, \ldots, v_n \geq 0$, weights $w_1, \ldots, w_n \geq 0$, and budget $B \geq 0$, we construct a max-potency instance as follows.

1. **Graph.** Let $\mathcal{G}$ be the disjoint union of n TP-cells $\mathcal{C}_1, \ldots, \mathcal{C}_n$. The observed variables are $\boldsymbol{V} = \{X_1, Y_1, \ldots, X_n, Y_n\}$.
2. **Observational distribution.** Choose per-cell distributions $P_i(X_i, Y_i)$ such that $pot_i(\tilde{a}_i) = v_i$, where $\tilde{a}_i = P(y_{i,x_i})$. Constructive achievability is established in Lemma E.3 below; if necessary, rescale the values v_i so that they lie in the achievable range. Set $P(\boldsymbol{V}) = \prod_{i=1}^{n} P_i(X_i, Y_i)$.
3. **Query.** Let
$$\theta = \sum_{i=1}^{n} \mathrm{PNS}_i = \sum_{i=1}^{n} P(y_{i,x_i}, y'_{i,x'_i}).$$
4. **Experiment set.** Let $\mathcal{A} = \{\tilde{a}_1, \ldots, \tilde{a}_n\}$ with $\tilde{a}_i = P(y_{i,x_i})$. For any $S \subseteq \{1, \ldots, n\}$, we write $a_S := \{\tilde{a}_i : i \in S\}$.
5. **Cost function and budget.** Set $c(a_S) = \sum_{i \in S} w_i$ for $S \subseteq \{1, \ldots, n\}$, and use the original budget B.

The construction is polynomial in n: each TP-cell is a constant-size object, and Lemma E.3 produces the required P_i in time polynomial in the bit-length of v_i.

E.3 Additivity of potency across cells

The key technical claim is that the potency of any subset a_S decomposes as a sum of per-cell potencies. This follows from the district structure of $\mathcal{G}$ via the machinery developed in Section 6.

Lemma E.2 (Additivity)**.** *For any $S \subseteq \{1, \ldots, n\}$,*

$$pot_\theta(a_S) \;=\; \sum_{i \in S} pot_i(\tilde{a}_i).$$

Proof. Each TP-cell $\mathcal{C}_i$ is a connected component of the bidirected graph and shares no variables with any other cell, so the districts of $\mathcal{G}$ are exactly $\mathcal{C}_1, \ldots, \mathcal{C}_n$. Let R_i denote the response-types of cell i. By Lemma 6.5 (Cartesian product decomposition), $h(\boldsymbol{v}) = \prod_{i=1}^{n} h_i(v_i)$ for any $\boldsymbol{v} = (v_1, \ldots, v_n) \in \mathrm{supp}(\boldsymbol{V})$, and response-types in distinct cells are independent: $P(\boldsymbol{r}) = \prod_{i=1}^{n} P(r_i)$.

We now show that the polynomial program $\mathcal{P}_{a_S}$ decomposes into n independent per-cell programs.

Objective separates. Each PNS_i involves only interventions, outcomes, and response-types of cell i, so the target polynomial $\mathcal{T} = \sum_i \mathrm{PNS}_i$ is a sum where the i-th summand depends only on R_i.

Observational constraints separate. The joint constraint $P(\boldsymbol{v}) = \sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(\boldsymbol{r})$ is implied by the per-cell constraints $P_i(v_i) = \sum_{r_i \in h_i(v_i)} P(r_i)$ for $i = 1, \ldots, n$, since $P(\boldsymbol{v}) = \prod_i P_i(v_i)$ and $h(\boldsymbol{v}) = \prod_i h_i(v_i)$.

Experimental constraints separate. Each constraint $f_{\tilde{a}_i}(\boldsymbol{R}) = p_{\tilde{a}_i}$ involves only R_i.

Hence $\mathcal{P}_{a_S}$ decomposes as

$$U_\theta(a_S, p_{a_S}) = \sum_{i \in S} U_i(\tilde{a}_i, p_{\tilde{a}_i}) + \sum_{i \notin S} U_i(\emptyset),$$

and similarly for L_θ, giving the instantiated width

$$W_\theta(a_S, p_{a_S}) = \sum_{i \in S} W_i(\tilde{a}_i, p_{\tilde{a}_i}) + \sum_{i \notin S} W_i(\emptyset).$$

The worst-case is taken over the product domain $p_{a_S} \in \prod_{i \in S} [l_{p_{\tilde{a}_i}}, u_{p_{\tilde{a}_i}}]$. Since the sum is separable in the $p_{\tilde{a}_i}$, the maximum decomposes:

$$\mathcal{W}_\theta(a_S) = \max_{p_{a_S}} W_\theta(a_S, p_{a_S}) = \sum_{i \in S} \mathcal{W}_i(\tilde{a}_i) + \sum_{i \notin S} W_i(\emptyset).$$

For $S = \emptyset$ this gives $\mathcal{W}_\theta(\emptyset) = \sum_i W_i(\emptyset)$. Combining,

$$\begin{aligned} pot_\theta(a_S) &= \mathcal{W}_\theta(\emptyset) - \mathcal{W}_\theta(a_S) \\ &= \sum_i W_i(\emptyset) - \sum_{i \in S} \mathcal{W}_i(\tilde{a}_i) - \sum_{i \notin S} W_i(\emptyset) \\ &= \sum_{i \in S} \big[W_i(\emptyset) - \mathcal{W}_i(\tilde{a}_i) \big] = \sum_{i \in S} pot_i(\tilde{a}_i). \end{aligned}$$

□

E.4 Achievability of target potencies

Lemma E.3 (Constructive achievability)**.** *There exists $v^* > 0$ and a polynomial-time algorithm that, given any rational $v \in [0, v^*]$, outputs rational parameters $(\alpha, \beta, \gamma, \delta) \in \Delta$ with $pot_i(\tilde{a}_i) = v$.*

Proof. We exhibit a one-parameter family of cells indexed directly by v. For $v \in [0,\ 1/8]$, define

$$P^v := \big(\alpha,\ \beta,\ \gamma,\ \delta\big) = \big(1 - 8v,\ v,\ 3v,\ 4v\big).$$

All four entries are non-negative on $[0,\ 1/8]$ and sum to 1, so $P^v \in \Delta$. We will show $pot_i(\tilde{a}_i;\ P^v) = v$ on this range, and set $v^* := 0.1$.

Observational width. By the Tian–Pearl observational bound stated above,

$$W_i(\emptyset;\ P^v) = \alpha + \delta = 1 - 4v. \tag{1}$$

Width with the experimental constraint. We now use the Tian–Pearl bounds for PNS when the interventional marginal $p = P(y_{i,x_i})$ is known. With the experimental constraint $\tilde{a}_i = P(y_{i,x_i}) = p$ added to the observational data, the bounds still involve the unobserved interventional marginal

$$q := P(y_{i,x_i'}).$$

Since q is not part of the allowed experiment set in the reduction, it is optimized out over its observationally feasible range

$$q \in [\gamma, 1 - \delta] = [3v, 1 - 4v].$$

Thus the sharp lower and upper bounds on PNS_i, conditional on observing p, are obtained as

$$\begin{aligned} L_i(\tilde{a}_i,\ p) &= \min_{q \in [3v,\ 1-4v]} \max\big\{0,\ p - q,\ P(y) - q,\ p - P(y)\big\}, \\ U_i(\tilde{a}_i,\ p) &= \max_{q \in [3v,\ 1-4v]} \min\big\{p,\ 1 - q,\ \alpha + \delta,\ p - q + \beta + \gamma\big\}. \end{aligned}$$

Substituting the marginals $P(y) = \alpha + \gamma = 1 - 5v$, $\alpha + \delta = 1 - 4v$, and $\beta + \gamma = 4v$:

Lower bound. The expression $\max\{0,\ p - q,\ (1 - 5v) - q,\ p - (1 - 5v)\}$ is non-increasing in q (each term is either independent of q or non-increasing in q), so the minimum over $q \in [3v,\ 1 - 4v]$ is attained at $q = 1 - 4v$. Substituting,

$$L_i(\tilde{a}_i,\ p) = \max\big\{0,\ p + 4v - 1,\ -v,\ p + 5v - 1\big\} = \max\big\{0,\ p + 5v - 1\big\},$$

where we drop $-v \le 0$ and note $p + 5v - 1 \ge p + 4v - 1$.

Upper bound. The expression $\min\{p,\ 1-q,\ 1-4v,\ p-q+4v\}$ is non-decreasing in $(-q)$, so the maximum over $q \in [3v,\ 1-4v]$ is attained at $q = 3v$. Substituting,

$$U_i(\tilde{\mathfrak{a}}_i,\ p) = \min\{p,\ 1-3v,\ 1-4v,\ p+v\} = \min\{p,\ 1-4v\},$$

since $1-4v \le 1-3v$ and $p \le p+v$.

The bound width $W_i(\tilde{\mathfrak{a}}_i,\ p;\ P^v) = U_i(\tilde{\mathfrak{a}}_i,\ p) - L_i(\tilde{\mathfrak{a}}_i,\ p)$ follows by case analysis on $p \in [1-8v,\ 1-v]$:

$$W_i(\tilde{\mathfrak{a}}_i,\ p;\ P^v) = \begin{cases} p, & p \in [\,1-8v,\ 1-5v\,], \\ 1-5v, & p \in [\,1-5v,\ 1-4v\,], \\ (2-9v)-p, & p \in [\,1-4v,\ 1-v\,]. \end{cases} \tag{2}$$

The function $p \mapsto W_i(\tilde{\mathfrak{a}}_i,\ p;\ P^v)$ is continuous and piecewise-linear: it rises with slope $+1$ from $W = 1-8v$ to $W = 1-5v$, plateaus at $1-5v$ on $[1-5v,\ 1-4v]$, then falls with slope -1 to $W = 1-8v$. The maximum is attained on the plateau, giving

$$\mathcal{W}_i(\tilde{\mathfrak{a}}_i;\ P^v) = \max_{p \in [1-8v,\ 1-v]} W_i(\tilde{\mathfrak{a}}_i,\ p;\ P^v) = 1-5v. \tag{3}$$

Potency. Combining (1) and (3),

$$pot_i(\tilde{\mathfrak{a}}_i;\ P^v) = W_i(\emptyset;\ P^v) - \mathcal{W}_i(\tilde{\mathfrak{a}}_i;\ P^v) = (1-4v) - (1-5v) = v.$$

At $v = 0.1$ this recovers $pot_i(\tilde{\mathfrak{a}}_i;\ P^{0.1}) = 0.1$ in agreement with Example 3.5.

Polynomial time. Given a rational v with bit-length L, the four entries $(1-8v,\ v,\ 3v,\ 4v)$ of P^v are computed by a constant number of rational arithmetic operations and have bit-length $\mathcal{O}(L)$. The cell is therefore constructed in time polynomial in the bit-length of v. □

E.5 The reduction

Theorem E.4 (Theorem 4.1, restated)**.** *The max-potency problem is NP-hard.*

Proof. We exhibit a polynomial-time reduction from 0-1 knapsack to max-potency. Let $(v_1, \ldots, v_n;\ w_1, \ldots, w_n;\ B)$ be a 0-1 knapsack instance.

Rescaling. If $\max_i v_i > v^*$, replace each v_i with $\tilde{v}_i := v_i \cdot v^* / \max_j v_j$, so that $\tilde{v}_i \in [0, v^*]$. Since this rescaling multiplies every subset sum $\sum_{i \in S} v_i$ by the same positive constant $v^* / \max_j v_j$, it preserves the argmax: the optimal S^* under $(\tilde{v}, w, B)$ is identical to the optimal S^* under (v, w, B). The rescaled values $\tilde{v}_i$ are rational with bit-length polynomial in the original input. Henceforth we work with $\tilde{v}_i$ in place of v_i.

Construction. Construct the max-potency instance $(\mathcal{G}, P(\boldsymbol{V}), \theta, \mathcal{A}, c, B)$ as in Section E.2, with cell distributions chosen via Lemma E.3 so that $pot_i(\tilde{\mathfrak{a}}_i) = \tilde{v}_i$. The construction runs in time polynomial in the size of the knapsack instance.

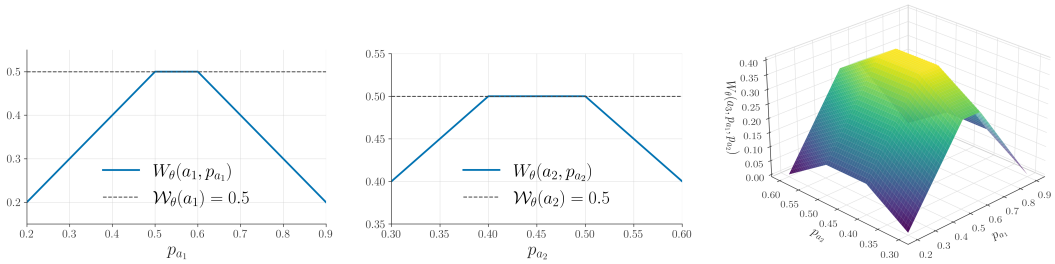
Equivalence. By Lemma E.2, for any $S \subseteq \{1, \ldots, n\}$,

$$pot_\theta(\mathfrak{a}_S) = \sum_{i \in S} pot_i(\tilde{\mathfrak{a}}_i) = \sum_{i \in S} \tilde{v}_i,$$

and by construction $c(\mathfrak{a}_S) = \sum_{i \in S} w_i$. The max-potency problem on the constructed instance is therefore

$$\max_{S \subseteq \{1,\ldots,n\}} \sum_{i \in S} \tilde{v}_i \quad \text{s.t.} \quad \sum_{i \in S} w_i \le B,$$

whose argmax coincides, by the rescaling argument, with the argmax of the original 0-1 knapsack instance. Since 0-1 knapsack is NP-hard, max-potency is NP-hard. □



(a) Experiment $\tilde{a}_1$. (b) Experiment $\tilde{a}_2$. (c) Experiment $a_3 = \{\tilde{a}_1, \tilde{a}_2\}$.

Figure 6: Example 3.5. The realized bound width $W_\theta(a, p_a) := U_\theta(a, p_a) - L_\theta(a, p_a)$ is shown as a function of p_a.

F Additional Examples

Example F.1 (Cartesian product decomposition)**.** *Consider the chain $A \to B \to C \to D$ with confounders R_1 on $\{A, B\}$ and R_2 on $\{C, D\}$ (Figure 7). Structural equations: $A = f_A(R_1)$, $B = f_B(A, R_1)$, $C = f_C(B, R_2)$, $D = f_D(C, R_2)$. The districts are $\boldsymbol{D}_1 = \{R_1, A, B\}$ and $\boldsymbol{D}_2 = \{R_2, C, D\}$ with response-types $\boldsymbol{R}_{\boldsymbol{D}_1} = \{R_1\}$ and $\boldsymbol{R}_{\boldsymbol{D}_2} = \{R_2\}$.*

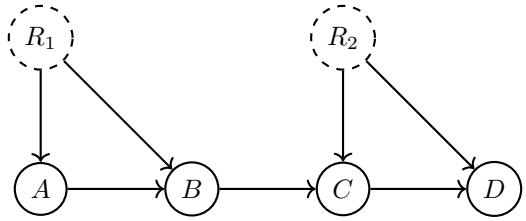


Figure 7: Chain graph with two districts and confounders R_1, R_2.

We label response-type values by the function they encode. For A (no parents), a_0 and a_1 denote the constants $A = 0$ and $A = 1$. For B with parent A, the four functions $\{0, 1\} \to \{0, 1\}$ are labeled following the response-type terminology of the literature [Angrist et al., 1996, Greenland and Robins, 1986]:

	b_{nev} (never-taker)	b_{com} (complier)	b_{def} (defier)	b_{alw} (always-taker)
$f_B(A{=}0, \cdot)$	0	0	1	1
$f_B(A{=}1, \cdot)$	0	1	0	1

The behaviors of C (w.r.t. parent B) and D (w.r.t. parent C) are labeled analogously. Fix $\boldsymbol{v} = (A{=}1, B{=}0, C{=}1, D{=}0)$:

$$h_1(\boldsymbol{v}) = \{r_1 : f_A(r_1) = 1 \ \wedge \ f_B(1, r_1) = 0\} = \{(a_1, b_{nev}),\ (a_1, b_{def})\},$$
$$h_2(\boldsymbol{v}) = \{r_2 : f_C(0, r_2) = 1 \ \wedge \ f_D(1, r_2) = 0\} = \{(c_{def}, d_{nev}),\ (c_{def}, d_{def}),\ (c_{alw}, d_{nev}),\ (c_{alw}, d_{def})\}.$$

By Lemma 6.5, $h(\boldsymbol{v}) = h_1(\boldsymbol{v}) \times h_2(\boldsymbol{v})$ with $|h(\boldsymbol{v})| = 2 \times 4 = 8$. For verification: $\boldsymbol{r} = (a_1, b_{nev}, c_{def}, d_{nev})$ yields $f_A(a_1) = 1$, $f_B(1, b_{nev}) = 0$, $f_C(0, c_{def}) = 1$, $f_D(1, d_{nev}) = 0$ — consistent with $\boldsymbol{v}$ and therefore $\boldsymbol{r} \in h(\boldsymbol{v})$.

Example F.2 (Step-by-step trace of Algorithm 4)**.** *We run* ALLPATHSINTERCEPTED *on Figure 1b with $\boldsymbol{R}^*_\theta = \{R_1, R_2\}$, $\boldsymbol{W} = \{D\}$, $\boldsymbol{Z} = \{C\}$. Removing $\boldsymbol{Z}$ yields the graph in Figure 8 (edges to/from C deleted).*

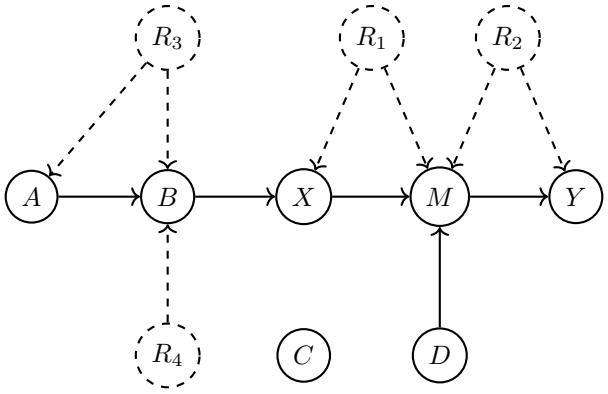


Figure 8: Figure 1b with all edges to/from C removed (i.e., after deleting $\boldsymbol{Z} = \{C\}$).

After initialization: $\texttt{visited} = \{R_1, R_2\}$, $\texttt{queue} = (R_1, R_2)$.

Iteration 1: $\texttt{current} = R_1$. *Not in* $\boldsymbol{W}$. *Children in* $\boldsymbol{E}'$*:* X, M *(unvisited). Update* $\texttt{visited} = \{R_1, R_2, X, M\}$, $\texttt{queue} = (R_2, X, M)$.

Iteration 2: $\texttt{current} = R_2$. *Unvisited child:* Y. *Update* $\texttt{visited} = \{R_1, R_2, X, M, Y\}$, $\texttt{queue} = (X, M, Y)$.

Iterations 3–5: X *has no unvisited children;* M*'s child* Y *already visited;* Y *has no children.*

Return `True`*. No path from* $\boldsymbol{R}^*_\theta$ *to* D *exists in* $\boldsymbol{E}'$*. By Corollary 6.13, this intervention is useless.*

G Algorithmic Details

G.1 Constructing the Polynomial Program

Canonicalization. We assume throughout that the input ADMG is canonicalized—no disturbance has a parent in $\mathcal{G}$, and no disturbance influences a strict subset of another's children. This is without loss of generality for the full data law [Duarte et al., 2024, Proposition 1]; see Appendix B.1 of that paper for the canonicalization procedure.

Constructing $\mathcal{P}_\emptyset$. The base program is obtained by invoking Algorithm 2 of Duarte et al. [2024] with inputs $(\mathcal{G}, P(\boldsymbol{V}), \emptyset, \mathit{supp}(\boldsymbol{V}), \theta)$.

Constructing $\mathcal{P}_a$. Given the base program, each experimental statement is polynomialized via the same procedure applied to the equality $\tilde{a} = p_{\tilde{a}}$:

Algorithm 2 Formulating $\mathcal{P}_a$

Require: Base Program $\mathcal{P}_\emptyset = (\mathcal{T}, \mathcal{C}_\emptyset)$, Experiments $a \subseteq \mathcal{A}$
Ensure: Program $\mathcal{P}_a = (\mathcal{T}, \mathcal{C}_a)$
$\mathcal{C}_a \leftarrow \mathcal{C}_\emptyset$ ▷ Initialize with base constraints
for $\tilde{a} \in a$ **do**
  $c_{\tilde{a}} \leftarrow$ duarte.polynomialize$(\tilde{a} = p_{\tilde{a}})$
  $\mathcal{C}_a \leftarrow \mathcal{C}_a \cup \{c_{\tilde{a}}\}$
end for
return $\mathcal{P}_a = (\mathcal{T}, \mathcal{C}_a)$

G.2 Evaluating Potency

For a fixed experiment set a, potency evaluation requires the worst-case post-experimental width

$$\mathcal{W}_\theta(a) = \max_{p_a} \left[U_\theta(a, p_a) - L_\theta(a, p_a) \right].$$

Unfolding the inner bounds as polynomial programs over independent optimizers P^U and P^L, both required to satisfy the observational constraints and to reproduce the experimental outcome p_a,

$$\mathcal{W}_\theta(a) = \max_{p_a} \max_{\substack{P^U \in \mathcal{C}_\emptyset \\ f(P^U)=p_a}} \max_{\substack{P^L \in \mathcal{C}_\emptyset \\ f(P^L)=p_a}} \big[\mathcal{T}(P^U) - \mathcal{T}(P^L)\big].$$

The variable p_a appears only as the common value of $f(P^U)$ and $f(P^L)$. We can therefore drop it as a decision variable and replace its two binding constraints with the single equality $f(P^U) = f(P^L)$. Whatever value the two optimizers agree on *is* the experimental outcome; no separate enumeration of p_a is needed.

This also resolves the question of feasibility for p_a. In the outer-loop view one would first compute bounds on p_a under $\mathcal{C}_\emptyset$ and grid over them; here those bounds are enforced implicitly, since the existence of $P^U, P^L \in \mathcal{C}_\emptyset$ producing a common value p is exactly the definition of p being feasible. See Algorithm 3.

Algorithm 3 EVALUATEPOTENCY

Require: Base program $\mathcal{P}_\emptyset = (\mathcal{T}, \mathcal{C}_\emptyset)$, experiment set a
Ensure: Potency $pot_\theta(a)$

1: $L_\theta \leftarrow \min \mathcal{T}$ subject to $\mathcal{C}_\emptyset$
2: $U_\theta \leftarrow \max \mathcal{T}$ subject to $\mathcal{C}_\emptyset$
3: $\mathcal{W}_\theta(\emptyset) \leftarrow U_\theta - L_\theta$
4: Create two copies of the response-type probabilities:

$$\{P^U(\boldsymbol{r}) : \boldsymbol{r} \in supp(\boldsymbol{R})\}, \qquad \{P^L(\boldsymbol{r}) : \boldsymbol{r} \in supp(\boldsymbol{R})\}.$$

5: Let $\mathcal{C}_\emptyset^U$ and $\mathcal{C}_\emptyset^L$ denote copies of $\mathcal{C}_\emptyset$ using $P^U(\boldsymbol{r})$ and $P^L(\boldsymbol{r})$, respectively.
6: Form the coupled constraint set

$$\mathcal{C}_a^{\text{cpl}} = \mathcal{C}_\emptyset^U \cup \mathcal{C}_\emptyset^L \cup \left\{ f_{\tilde{a}}(\boldsymbol{R}^U) = f_{\tilde{a}}(\boldsymbol{R}^L) : \tilde{a} \in a \right\}.$$

7: Compute

$$\mathcal{W}_\theta(a) \leftarrow \max \left[\mathcal{T}(\boldsymbol{R}^U) - \mathcal{T}(\boldsymbol{R}^L)\right] \quad \text{s.t.} \quad \mathcal{C}_a^{\text{cpl}}.$$

8: **return** $\mathcal{W}_\theta(\emptyset) - \mathcal{W}_\theta(a)$

The formulation is exact at the mathematical level; numerical solutions are subject to the tolerances of the polynomial optimizer.

G.3 AllPathsIntercepted Algorithm

Algorithm 4 AllPathsIntercepted

Require: Graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{R}, \boldsymbol{E})$, $\boldsymbol{R}_\theta^*$, Outcomes $\boldsymbol{W}$, Interventions $\boldsymbol{Z}$
Ensure: `True` iff every directed path $\boldsymbol{R}_\theta^* \to \boldsymbol{W}$ passes through $\boldsymbol{Z}$

$\boldsymbol{E}' \leftarrow \{(u,v) \in \boldsymbol{E} : u \notin \boldsymbol{Z} \wedge v \notin \boldsymbol{Z}\}$ ▷ Remove edges involving $\boldsymbol{Z}$
`visited` $\leftarrow \boldsymbol{R}_\theta^*$; `queue` $\leftarrow$ `list`$(\boldsymbol{R}_\theta^*)$
while `queue` $\neq \emptyset$ **do**
  `current` $\leftarrow$ `queue.pop()`
  if `current` $\in \boldsymbol{W}$ **then return** `False` ▷ Unintercepted path found
  end if
  for $c \in$ `children(current,` $\boldsymbol{E}'$`)` with $c \notin$ `visited` **do**
    `visited` $\leftarrow$ `visited` $\cup \{c\}$; `queue.append(`c`)`
  end for
end while
return `True` ▷ All paths intercepted

Runtime of Algorithm 4

We analyze each component of Algorithm 4:

Initialization: Constructing $\boldsymbol{E}'$ requires iterating through all edges and performing two membership checks per edge, yielding $\mathcal{O}(|\boldsymbol{E}|)$. Initializing `visited` and `queue` with $\boldsymbol{R}_\theta^*$ takes $\mathcal{O}(|\boldsymbol{R}_\theta^*|) \leq \mathcal{O}(|\boldsymbol{V}|)$ (this inequality holds in canonical DAGs).

BFS Loop: Each node in $\boldsymbol{V} \setminus \boldsymbol{Z}$ is added to `visited` at most once and consequently to `queue` at most once. Therefore, the while-loop executes at most $|\boldsymbol{V} \setminus \boldsymbol{Z}| \leq |\boldsymbol{V}|$ iterations, contributing $\mathcal{O}(|\boldsymbol{V}|)$ for dequeue operations and membership checks.

For the inner loop, each edge $(u, v) \in \boldsymbol{E}'$ is examined at most once—when node u is dequeued and we iterate over its children. Since every edge is examined at most once across the entire execution, the total work by the inner loop is $\mathcal{O}(|\boldsymbol{E}'|) \subseteq \mathcal{O}(|\boldsymbol{E}|)$.

Total:

$$\underbrace{\mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|)}_{\text{init.}} + \underbrace{\mathcal{O}(|\boldsymbol{V}|)}_{\text{while-loop}} + \underbrace{\mathcal{O}(|\boldsymbol{E}|)}_{\text{edge exams.}} = \mathcal{O}(|\boldsymbol{V}| + |\boldsymbol{E}|).$$

G.4 Solving the Max-Potency Problem

We use EVALUATEPOTENCY as a subroutine to solve the budget-constrained max-potency problem exactly. Since the problem is NP-hard (Appendix E), the algorithm is exponential in the worst case.

Algorithm 1 certifies a set $\mathcal{A}^\dagger$ of individually useless experiments. Let $\mathcal{A}^\dagger_{\text{ID}} \subseteq \mathcal{A}^\dagger$ denote those pruned by the ID check and $\mathcal{A}^\dagger_{\text{Int}} = \mathcal{A}^\dagger \setminus \mathcal{A}^\dagger_{\text{ID}}$ those pruned by interception. By Corollary 6.16, $\mathcal{A}^\dagger_{\text{ID}}$ can be removed from the search entirely. The algorithm then enumerates all subsets of $\mathcal{A} \setminus \mathcal{A}^\dagger_{\text{ID}}$, skipping singletons in $\mathcal{A}^\dagger_{\text{Int}}$.

Algorithm 5 SOLVEMAXPOTENCY

Require: Graph $\mathcal{G}$, query θ, base program $\mathcal{P}_\emptyset = (\mathcal{T}, \mathcal{C}_\emptyset)$, candidate experiments $\mathcal{A}$, cost function $c : 2^{\mathcal{A}} \to \mathbb{R}_+$, budget B
Ensure: Optimal experiment set $a^\star$ and potency value $v^\star$

$\mathcal{A}^\dagger \leftarrow$ GETUSELESSEXPERIMENTS$(\mathcal{G}, \theta, \mathcal{A})$
Partition $\mathcal{A}^\dagger$ into $\mathcal{A}^\dagger_{\text{ID}}$ and $\mathcal{A}^\dagger_{\text{Int}}$
$a^\star \leftarrow \emptyset$; $v^\star \leftarrow 0$
for $a \subseteq \mathcal{A} \setminus \mathcal{A}^\dagger_{\text{ID}}$ **do**
    if $|a| = 1$ **and** $a \subseteq \mathcal{A}^\dagger_{\text{Int}}$ **then**
        continue
    end if
    if $c(a) > B$ **then**
        continue
    end if
    $v \leftarrow$ EVALUATEPOTENCY$(\mathcal{P}_\emptyset, a)$
    if $v > v^\star$ **then**
        $v^\star \leftarrow v$
        $a^\star \leftarrow a$
    end if
end for
return $a^\star, v^\star$

The algorithm is exact provided that EVALUATEPOTENCY (Algorithm 3) globally solves the coupled polynomial program. Its runtime is $\mathcal{O}(2^{|\mathcal{A}|-|\mathcal{A}^\dagger_{\text{ID}}|} \cdot T_{\text{pot}})$, where T_{pot} denotes the cost of one potency evaluation. Corollary 6.16 reduces the exponent from $|\mathcal{A}|$ to $|\mathcal{A}| - |\mathcal{A}^\dagger_{\text{ID}}|$.

H Omitted Proofs

Feasible sets. We define two types of feasible sets used in the proofs below. The *observational feasible set* $\mathcal{F}_\emptyset := \{\boldsymbol{r} \in supp(\boldsymbol{R}) : \text{all observational constraints from } P(\boldsymbol{V}) \text{ are satisfied}\}$ is the feasible set of the base program $\mathcal{P}_\emptyset$ as constructed by Duarte et al. [2024]. For an experiment $\boldsymbol{a}$ with observed outcome $p_{\boldsymbol{a}}$, the *experimental feasible set* is $\mathcal{F}_{\boldsymbol{a}}(p_{\boldsymbol{a}}) := \{\boldsymbol{r} \in supp(\boldsymbol{R}) : f_{\boldsymbol{a}}(\boldsymbol{r}) = p_{\boldsymbol{a}}\}$. The combined feasible set is $\mathcal{F}_\emptyset \cap \mathcal{F}_{\boldsymbol{a}}(p_{\boldsymbol{a}})$.

H.1 Proof of Lemma 3.3

Proof. Let $p_{\boldsymbol{a}} \in \prod_{\tilde{\boldsymbol{a}} \in \boldsymbol{a}} [l_{p_{\tilde{\boldsymbol{a}}}}, u_{p_{\tilde{\boldsymbol{a}}}}]$. The combined feasible set $\mathcal{F}_\emptyset \cap \mathcal{F}_{\boldsymbol{a}}(p_{\boldsymbol{a}}) \subseteq \mathcal{F}_\emptyset$ by definition of intersection. From optimization theory, restricting to a smaller feasible set can only decrease the maximum and increase the minimum. Hence $U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) \leq U_\theta$ and $L_\theta \leq L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}})$. It follows $\forall p_{\boldsymbol{a}}$:

$$0 \leq U_\theta - U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) \ \wedge \ L_\theta - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) \leq 0 \implies L_\theta - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) \leq U_\theta - U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}})$$
$$\implies U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) \leq U_\theta - L_\theta = \mathcal{W}_\theta$$

Since this holds for all $p_{\boldsymbol{a}}$, we have $\mathcal{W}_\theta(\boldsymbol{a}) = \max_{p_{\boldsymbol{a}}} \big(U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}})\big) \leq \mathcal{W}_\theta$, and hence $pot_\theta(\boldsymbol{a}) \geq 0$. □

H.2 Proof of Lemma 3.4

Proof. **(⇒)** Unpacking the definition yields

$$pot_\theta(\boldsymbol{a}) = 0 \implies \mathcal{W}_\theta(\boldsymbol{a}) = \mathcal{W}_\theta \implies \max_{p_{\boldsymbol{a}} \in \prod_{\tilde{\boldsymbol{a}} \in \boldsymbol{a}} [l_{p_{\tilde{\boldsymbol{a}}}}, u_{p_{\tilde{\boldsymbol{a}}}}]} \big(U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}})\big) = U_\theta - L_\theta.$$

Let $p^\dagger_{\boldsymbol{a}} = \underset{p_{\boldsymbol{a}} \in \prod_{\tilde{\boldsymbol{a}} \in \boldsymbol{a}} [l_{p_{\tilde{\boldsymbol{a}}}}, u_{p_{\tilde{\boldsymbol{a}}}}]}{\arg\max} U_\theta(\boldsymbol{a}, p_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p_{\boldsymbol{a}})$:

$$U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = U_\theta - L_\theta \implies U_\theta - U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = L_\theta - L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}).$$

As we saw in the proof of Lemma 3.3, $L_\theta \leq L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}})$ and $U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) \leq U_\theta$, which implies that[6]

$$U_\theta - U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = 0 \ \wedge \ L_\theta - L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = 0 \implies U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = U_\theta \ \wedge \ L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = L_\theta.$$

(⇐) Suppose there exists $p^\dagger_{\boldsymbol{a}} \in \prod_{\tilde{\boldsymbol{a}} \in \boldsymbol{a}} [l_{p_{\tilde{\boldsymbol{a}}}}, u_{p_{\tilde{\boldsymbol{a}}}}]$ with $U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = U_\theta$ and $L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = L_\theta$. Then

$$U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) - L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) = U_\theta - L_\theta = \mathcal{W}_\theta.$$

Since $\mathcal{W}_\theta(\boldsymbol{a})$ is the maximum over all $p_{\boldsymbol{a}}$, it must hold that $\mathcal{W}_\theta(\boldsymbol{a}) \geq \mathcal{W}_\theta$. By Lemma 3.3, $\mathcal{W}_\theta(\boldsymbol{a}) \leq \mathcal{W}_\theta$, hence $\mathcal{W}_\theta(\boldsymbol{a}) = \mathcal{W}_\theta$ and $pot_\theta(\boldsymbol{a}) = 0$. □

H.3 Proof of Lemma 6.5 (Cartesian Product Decomposition)

Proof. We start by applying Definition 6.1 to $h(\boldsymbol{v})$, simplified for the case of $|\mathfrak{C}| = 1$:

$$h(\boldsymbol{v}) = \{\boldsymbol{r} : \forall V \in \boldsymbol{V}, F_V(\boldsymbol{r}) = v\}.$$

Since $\boldsymbol{X} = \emptyset$ (no interventions), $F_V(\boldsymbol{r})$ reduces to the structural equation with all parent values recursively fixed by $\boldsymbol{v}$:

$$F_V(\boldsymbol{r}) = f_V\left(\bigwedge_{P \in \boldsymbol{pa}(V)} \underbrace{f(f(\ldots f(\boldsymbol{r}), \boldsymbol{r}), \boldsymbol{r})}_{\text{equals elements of } \boldsymbol{v}}, \boldsymbol{r}\right).$$

Using Definition 6.3 to replace the recursive evaluation:

$$h(\boldsymbol{v}) = \{\boldsymbol{r} : \forall V \in \boldsymbol{V}, \ f_V(\boldsymbol{pa}_{\boldsymbol{v}}, \boldsymbol{r}) = v\}.$$

[6] Since $U_\theta - U_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) \geq 0$ and $L_\theta - L_\theta(\boldsymbol{a}, p^\dagger_{\boldsymbol{a}}) \leq 0$, their equality forces both to be zero.

Rewriting $\forall V \in \boldsymbol{V}$ as $\forall \boldsymbol{D}_k \in \boldsymbol{D}, \forall V \in \boldsymbol{D}_k \cap \boldsymbol{V}$ and partitioning $\boldsymbol{r}$ into per-district vectors $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{|\boldsymbol{D}|}$:

$$h(\boldsymbol{v}) = \{(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{|\boldsymbol{D}|}) : \forall \boldsymbol{D}_k \in \boldsymbol{D}, \forall V \in \boldsymbol{D}_k \cap \boldsymbol{V},\ f_V(\boldsymbol{pa}_{\boldsymbol{v}}, \boldsymbol{r}) = v\}.$$

Since $f_V(\boldsymbol{pa}_{\boldsymbol{v}}, \boldsymbol{r}) = f_V(\boldsymbol{pa}_{\boldsymbol{v}}, \boldsymbol{r}_k)$ whenever $V \in \boldsymbol{D}_k$:

$$h(\boldsymbol{v}) = \{(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{|\boldsymbol{D}|}) : \forall \boldsymbol{D}_k \in \boldsymbol{D}, \forall V \in \boldsymbol{D}_k \cap \boldsymbol{V},\ f_V(\boldsymbol{pa}_{\boldsymbol{v}}, \boldsymbol{r}_k) = v\}.$$

The predicate is a conjunction where each conjunct depends only on the response-type configuration $\boldsymbol{r}_k$ of one district $\boldsymbol{D}_k$. By Lemma I.1 (Appendix I), we can factorize this set:

$$h(\boldsymbol{v}) = \prod_{\boldsymbol{D}_k \in \boldsymbol{D}} \{\boldsymbol{r}_k : \forall V \in \boldsymbol{V} \cap \boldsymbol{D}_k,\ f_V(\boldsymbol{pa}_{\boldsymbol{v}}(V), \boldsymbol{r}_k) = v\} = \prod_{\boldsymbol{D}_k \in \boldsymbol{D}} h_k(\boldsymbol{v}).$$

□

H.4 Proof of Lemma 6.7

Proof. **Single district.** By Eq. (26) from Tian and Pearl [2002], the c-factor for district $\boldsymbol{D}_k$ is

$$Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{u}_k \in supp(\boldsymbol{U}_k)} \prod_{V_i \in \boldsymbol{V}_k} P(v_i \mid \boldsymbol{pa}_{\boldsymbol{v}}(V_i), \boldsymbol{u}_k)\, P(\boldsymbol{u}_k),$$

where $\boldsymbol{U}_k$ denotes the latent confounders in $\boldsymbol{D}_k$. Replacing $\boldsymbol{U}_k$ by our discrete $\boldsymbol{R}_k$:

$$Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}_k} \prod_{V_i \in \boldsymbol{V}_k} P(v_i \mid \boldsymbol{pa}_{\boldsymbol{v}}(V_i), \boldsymbol{r}_k)\, P(\boldsymbol{r}_k).$$

Each realization $\boldsymbol{r}_k$ deterministically fixes the structural equations, so the conditional probability becomes an indicator:

$$Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}_k} \prod_{V_i \in \boldsymbol{V}_k} \mathbf{1}\big[f_{V_i}(\boldsymbol{pa}_{\boldsymbol{v}}(V_i),\, \boldsymbol{r}_k) = v_i\big]\, P(\boldsymbol{r}_k).$$

The product of indicators equals 1 if and only if $\boldsymbol{r}_k \in h_k(\boldsymbol{v})$, yielding $Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}_k \in h_k(\boldsymbol{v})} P(\boldsymbol{r}_k)$.

Set of districts. We show the result for $|\boldsymbol{D}'| = 2$; the general case follows by induction. Let $\boldsymbol{D}' = \{\boldsymbol{D}_1, \boldsymbol{D}_2\}$. Applying the single-district result:

$$\prod_{\boldsymbol{D}_k \in \boldsymbol{D}'} Q_k(\boldsymbol{v}) = \left(\sum_{r_1 \in h_1(\boldsymbol{v})} P(r_1) \right) \left(\sum_{r_2 \in h_2(\boldsymbol{v})} P(r_2) \right).$$

Expanding the product of sums:

$$= \sum_{r_1 \in h_1(\boldsymbol{v})} \sum_{r_2 \in h_2(\boldsymbol{v})} P(r_1)\, P(r_2).$$

Since response-types in distinct districts are independent, $P(r_{k_1})\, P(r_{k_2}) = P(\boldsymbol{r}_{\boldsymbol{D}'})$. Collapsing the double sum into a sum over the Cartesian product[7] and applying Lemma 6.5:

$$= \sum_{\boldsymbol{r}_{\boldsymbol{D}'} \in h_1(\boldsymbol{v}) \times h_2(\boldsymbol{v})} P(\boldsymbol{r}_{\boldsymbol{D}'}) = \sum_{\boldsymbol{r}_{\boldsymbol{D}'} \in h_{\boldsymbol{D}'}(\boldsymbol{v})} P(\boldsymbol{r}_{\boldsymbol{D}'}). \quad \square$$

H.5 Proof of Lemma 6.10

Proof. We write the observational constraint for a fixed $\boldsymbol{v} \in supp(\boldsymbol{V})$:

$$P(\boldsymbol{v}) = \sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(\boldsymbol{r}).$$

[7] Formally justified by Lemma I.2.

Let $\boldsymbol{D}_{\bar{\theta}} = \boldsymbol{D} \setminus \boldsymbol{D}^*_\theta$ denote all districts not in $\boldsymbol{D}^*_\theta$ and $\boldsymbol{R}_{\bar{\theta}} = \boldsymbol{D}_{\bar{\theta}} \cap \boldsymbol{R}$ their response types. Since response-types in distinct districts are independent, $P(\boldsymbol{r}) = P(\boldsymbol{r}^*_\theta)\, P(\boldsymbol{r}_{\bar{\theta}})$:

$$P(\boldsymbol{v}) = \sum_{\boldsymbol{r} \in h(\boldsymbol{v})} P(\boldsymbol{r}^*_\theta)\, P(\boldsymbol{r}_{\bar{\theta}}).$$

By Lemma 6.5, $h(\boldsymbol{v}) = \prod_{\boldsymbol{D}_k \in \boldsymbol{D}} h_k(\boldsymbol{v})$, which we partition as $h_{\boldsymbol{D}^*_\theta}(\boldsymbol{v}) \times h_{\boldsymbol{D}_{\bar{\theta}}}(\boldsymbol{v})$. Writing as a double summation and pulling out $P(\boldsymbol{r}^*_\theta)$ (independent of $\boldsymbol{r}_{\bar{\theta}}$):

$$P(\boldsymbol{v}) = \sum_{\boldsymbol{r}^*_\theta \in h_{\boldsymbol{D}^*_\theta}(\boldsymbol{v})} P(\boldsymbol{r}^*_\theta) \sum_{\boldsymbol{r}_{\bar{\theta}} \in h_{\boldsymbol{D}_{\bar{\theta}}}(\boldsymbol{v})} P(\boldsymbol{r}_{\bar{\theta}}).$$

Since $\boldsymbol{D}^*_\theta \cup \boldsymbol{D}_{\bar{\theta}} = \boldsymbol{D}$, the c-factor decomposition gives $P(\boldsymbol{v}) = \prod_{\boldsymbol{D}_k \in \boldsymbol{D}^*_\theta} Q_k(\boldsymbol{v}) \cdot \prod_{\boldsymbol{D}_l \in \boldsymbol{D}_{\bar{\theta}}} Q_l(\boldsymbol{v})$:

$$\prod_{\boldsymbol{D}_k \in \boldsymbol{D}^*_\theta} Q_k(\boldsymbol{v}) \prod_{\boldsymbol{D}_l \in \boldsymbol{D}_{\bar{\theta}}} Q_l(\boldsymbol{v}) = \sum_{\boldsymbol{r}^*_\theta \in h_{\boldsymbol{D}^*_\theta}(\boldsymbol{v})} P(\boldsymbol{r}^*_\theta) \sum_{\boldsymbol{r}_{\bar{\theta}} \in h_{\boldsymbol{D}_{\bar{\theta}}}(\boldsymbol{v})} P(\boldsymbol{r}_{\bar{\theta}}).$$

By Lemma 6.7, $\prod_{\boldsymbol{D}_l \in \boldsymbol{D}_{\bar{\theta}}} Q_l(\boldsymbol{v}) = \sum_{\boldsymbol{r}_{\bar{\theta}} \in h_{\boldsymbol{D}_{\bar{\theta}}}(\boldsymbol{v})} P(\boldsymbol{r}_{\bar{\theta}})$. Dividing both sides by $\prod_{\boldsymbol{D}_l \in \boldsymbol{D}_{\bar{\theta}}} Q_l(\boldsymbol{v}) > 0$ yields the desired identity:

$$\prod_{\boldsymbol{D}_k \in \boldsymbol{D}^*_\theta} Q_k(\boldsymbol{v}) = \sum_{\boldsymbol{r}^*_\theta \in h_{\boldsymbol{D}^*_\theta}(\boldsymbol{v})} P(\boldsymbol{r}^*_\theta).$$

□

Lemma H.1. *The base program $\mathcal{P}_\emptyset$ optimizes only over $\boldsymbol{R}^*_\theta$.*

Proof. First, by construction of $\boldsymbol{R}^*_\theta$, it covers all response-types necessary for expressing θ. Therefore the target polynomial $\mathcal{T}$ is only a function of $\boldsymbol{R}^*_\theta$. Second, Lemma 6.10 shows that all observational constraints depend only on $\boldsymbol{R}^*_\theta$. Hence $\mathcal{C}_\emptyset$ is only a function of $\boldsymbol{R}^*_\theta$. Since both the target function and the constraints depend only on $\boldsymbol{R}^*_\theta$, it is sufficient to optimize over $\boldsymbol{R}^*_\theta$ in $\mathcal{P}_\emptyset$. □

H.6 Proof of Theorem 6.12

Proof. Since $\mathcal{P}_\emptyset$ optimizes over $\boldsymbol{R}^*_\theta$ only (Lemma H.1), there is no sequence of co-occurring decision variables connecting $R^\star(\tilde{a})$ to $\mathcal{T}$. By Duarte et al. [2024], the constraint for $\tilde{a}$ can be dropped, resulting in $\mathcal{P}_{\tilde{a}} = \mathcal{P}_\emptyset$ and $pot_\theta(\tilde{a}) = 0$. □

H.7 Proof of Corollary 6.13

Proof. By assumption, for every $\mathfrak{c} \in \mathfrak{C}$, every directed path from $\boldsymbol{R}^*_\theta$ to $\boldsymbol{Y}^{(\mathfrak{c})}$ passes through $\boldsymbol{X}^{(\mathfrak{c})}$. By the definition of $R^\star(\tilde{a})$, a response-type R is in $R^\star(\tilde{a})$ only if there exists some $\mathfrak{c} \in \mathfrak{C}$ and a directed path from R to $\boldsymbol{Y}^{(\mathfrak{c})}$ that does *not* pass through $\boldsymbol{X}^{(\mathfrak{c})}$. Since no such path exists for any $\mathfrak{c}$, no element of $\boldsymbol{R}^*_\theta$ can be in $R^\star(\tilde{a})$. Hence $R^\star(\tilde{a}) \cap \boldsymbol{R}^*_\theta = \emptyset$. By Theorem 6.12, we have $pot_\theta(\tilde{a}) = 0$. □

H.8 Proof of Theorem 6.15

Proof. Since $\tilde{a}$ is ID, it is uniquely computable from $P(\boldsymbol{V})$. Therefore, we can determine a unique value $p^*_{\tilde{a}}$ for the experimental outcome, which satisfies $\forall \boldsymbol{r} \in \mathcal{F}_\emptyset : f_{\tilde{a}}(\boldsymbol{r}) = p^*_{\tilde{a}}$. Hence $\mathcal{F}_\emptyset \subseteq \mathcal{F}_{\tilde{a}}(p^*_{\tilde{a}})$, which gives:

$$\mathcal{F}_\emptyset \cap \mathcal{F}_{\tilde{a}}(p^*_{\tilde{a}}) = \mathcal{F}_\emptyset.$$

Since the combined feasible set coincides with $\mathcal{F}_\emptyset$, the optimization is unchanged: $\mathcal{W}_\theta(\tilde{a}) = \mathcal{W}_\theta(\emptyset)$ and $pot_\theta(\tilde{a}) = 0$. □

H.9 Proof of Corollary 6.16

Proof. Since $\tilde{a}$ is identifiable, it has a unique feasible outcome $p^*_{\tilde{a}}$ with $\mathcal{F}_\emptyset \cap \mathcal{F}_{\tilde{a}}(p^*_{\tilde{a}}) = \mathcal{F}_\emptyset$ (Theorem 6.15). For any $b \subseteq \mathcal{A}$ and feasible p_b:

$$\mathcal{F}_\emptyset \cap \mathcal{F}_{\{\tilde{a}\} \cup b}(p^*_{\tilde{a}}, p_b) = \mathcal{F}_\emptyset \cap \mathcal{F}_{\tilde{a}}(p^*_{\tilde{a}}) \cap \mathcal{F}_b(p_b) = \mathcal{F}_\emptyset \cap \mathcal{F}_b(p_b).$$

Since $p^*_{\tilde{a}}$ is the only feasible value of $p_{\tilde{a}}$, the worst-case width satisfies $\mathcal{W}_\theta(\{\tilde{a}\} \cup b) = \mathcal{W}_\theta(b)$, and hence $pot_\theta(\{\tilde{a}\} \cup b) = pot_\theta(b)$. □

I Helper Lemmas

Lemma I.1 (Factorization of Cartesian Products)**.** *Let $\boldsymbol{R}_1, \boldsymbol{R}_2$ be two sets of random variables with disjoint supports, and let ϕ_1, ϕ_2 be predicates depending only on configurations of $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$, respectively. Then*

$$\Big\{(\boldsymbol{r}_1, \boldsymbol{r}_2) \in \mathit{supp}(\boldsymbol{R}_1) \times \mathit{supp}(\boldsymbol{R}_2) : \phi_1(\boldsymbol{r}_1) \wedge \phi_2(\boldsymbol{r}_2)\Big\}$$
$$= \big\{\boldsymbol{r}_1 \in \mathit{supp}(\boldsymbol{R}_1) : \phi_1(\boldsymbol{r}_1)\big\} \times \big\{\boldsymbol{r}_2 \in \mathit{supp}(\boldsymbol{R}_2) : \phi_2(\boldsymbol{r}_2)\big\}.$$

Proof. ($\subseteq$): If $(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is in the left-hand side, then $\phi_1(\boldsymbol{r}_1)$ holds, so $\boldsymbol{r}_1$ is in the first factor, and $\phi_2(\boldsymbol{r}_2)$ holds, so $\boldsymbol{r}_2$ is in the second factor.

($\supseteq$): If $\boldsymbol{r}_1$ is in the first factor and $\boldsymbol{r}_2$ in the second, then both $\phi_1(\boldsymbol{r}_1)$ and $\phi_2(\boldsymbol{r}_2)$ hold, so $(\boldsymbol{r}_1, \boldsymbol{r}_2)$ is in the left-hand side.

The generalization to any finite collection $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_K$ follows by induction. □

Lemma I.2 (Summation over Cartesian Product as Double Summation)**.** *Let $\mathbb{A}$ be one of $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$. Let S, T be finite sets and $f\colon S \times T \to \mathbb{A}$. Then*

$$\sum_{(s,t)\, \in\, S \times T} f(s,t) = \sum_{s \in S} \sum_{t \in T} f(s,t).$$

Proof. We proceed by induction on $|T|$.

Base case. If $|T| = 0$, then $T = \emptyset$, so $S \times T = \emptyset$. Both sides equal 0 by the convention for empty summation.

Inductive step. Suppose the identity holds whenever $|T| = n \geq 0$, and let $|T| = n + 1$. Pick $t_0 \in T$, set $T' := T \setminus \{t_0\}$. Then $S \times T = (S \times T') \cup (S \times \{t_0\})$ (disjoint). By the sum over disjoint unions:

$$\sum_{(s,t) \in S \times T} f(s,t) = \sum_{(s,t) \in S \times T'} f(s,t) + \sum_{(s,t) \in S \times \{t_0\}} f(s,t).$$

By the induction hypothesis, the first summand equals $\sum_{s \in S} \sum_{t \in T'} f(s,t)$. The second equals $\sum_{s \in S} f(s, t_0)$. Combining via linearity of summation:

$$= \sum_{s \in S} \left(\sum_{t \in T'} f(s,t) + f(s,t_0) \right) = \sum_{s \in S} \sum_{t \in T} f(s,t). \quad \square$$

J Runtime and Scalability Experiments

This appendix quantifies the computational profile of the pipeline. To state the scope precisely: we do not claim a scalable solver for partial identification. Potency evaluation relies on solving the underlying response-type programs and necessarily inherits their computational cost. Our contributions are instead (i) the formulation of experiment selection under partial identification and its hardness (Theorem 4.1), (ii) an exact potency evaluation procedure wherever the underlying programs are tractable, and (iii) sound graph-based certificates that reduce how many expensive programs must be solved. Pruning only inspects the graph, so we evaluate it on graphs with up to 76 variables in Section 7. Measuring the solver time it *saves*, by contrast, means actually running the solver, so the experiments below are limited to small instances—those on which the solver terminates at all, using the single consumer laptop of Appendix K.6. All runtimes were measured on that machine (single-threaded), using the `autobound` library and the potency procedure of Appendix L.2. The experiments in this appendix impose tighter per-solve time limits than the NHANES potencies—60 seconds for the solvability check below, 600 seconds for each candidate evaluation—since a runtime measurement has to terminate, and here the limits do bind.

J.1 Scalability of Exact Potency Evaluation

To probe how far the response-type programs themselves scale, we sampled random instances (Erdős–Rényi graphs, two-world counterfactual queries, model-consistent random $P(\boldsymbol{V})$) and checked whether the observational bounds alone are solvable within a 60-second solver time limit: at $|\boldsymbol{V}| = 5$, 12 of 25 instances were solved (median 0.9s); at $|\boldsymbol{V}| = 6$, 3 of 25 (median 5.1s); at $|\boldsymbol{V}| = 7$, 1 of 25 (solved in 1.3s).[8] Solving times appear strongly bimodal. One structural observation might explain this: by Lemma 6.10, the size of the optimization program is governed not by $|\boldsymbol{V}|$ but by the district hull of the *query*, so even a larger graph can lead to a small program when the query has a localized district hull.[9] In general, the max-potency problem and polynomial programming remain NP-hard and thus computationally challenging.

J.2 Runtime With and Without Pruning

Selecting a single experiment. On a benchmark of 26 randomly generated graphs with $|\boldsymbol{V}| = 5$, drawn with the same generation protocol as above at confounding densities $c \in \{0.5, 1.0\}$ (sampled until 26 had observational bounds that solved within 60 seconds, as above), using all single-variable interventions supported by each graph (2–12 candidates per graph, 152 in total), evaluating every candidate took at least 26,753 seconds (7.4 hours; each candidate was allowed a 600-second timeout), whereas evaluating only the survivors took 3,849 seconds (1.1 hours)—a speedup of at least $6.95\times$. The pruning pass itself took about 6 milliseconds per instance.

The saving is larger than the drop in candidate count (152 to 48) would suggest, because the candidates the rules discard are also the slowest to evaluate. Table 5 shows the split. A candidate that survives pruning is solved in 2.5 seconds at the median, against 612 seconds for a candidate discarded by the interception rule. Of the 50 candidates discarded by that rule, 29 did not finish within the timeout; these contribute the time actually spent on them,[10] so that group's total and the $6.95\times$ are both lower bounds. Setting aside the 39 candidates that did not finish, the remaining 113 take 2,240 seconds in total, against 204 seconds for the 42 of them that pruning keeps—an $11\times$ saving.

We suspect the two groups differ because a discarded candidate is one whose experiment cannot change the bounds: its constraint is never active, leaving the objective flat across a large part of the search space, so the solver has little to guide it and struggles to confirm optimality even though the answer is known in advance to be zero. Pruning removes exactly these cases in our Erdős–Rényi graphs.

Table 5: Time to evaluate a single candidate on the $|\boldsymbol{V}| = 5$ benchmark, grouped by what the pruning rules do with it. Candidates that did not finish within the 600-second timeout contribute their measured wall-clock time, so the means and totals are under-estimates.

Candidate group	count	median	mean	unfinished	total
Kept (must be evaluated)	48	2.5 s	80 s	6	3,849 s
Discarded by the ID check	54	4.7 s	49 s	4	2,659 s
Discarded by interception	50	612 s	405 s	29	20,245 s
All candidates	152	5.4 s	176 s	39	26,753 s

Subset search. On a randomly generated graph from the same generator ($|\boldsymbol{V}| = 5$, $c = 1.0$) with 8 experiments, unit costs, and total budget $B = 2$, evaluating the potency of all affordable sets required 105.7 seconds and 36 programs; removing the candidates pruned by the ID criterion beforehand reduced this to 62.7 seconds and 21 programs, while yielding exactly the same optimum, as guaranteed by Corollary 6.16. Every program here finished, so both figures are exact. This is an illustration of the resulting saving on an instance small enough to enumerate exhaustively, not a scaling study.

[8] One instance at $|\boldsymbol{V}| = 7$ is counted as unsolved for a different reason: building its program exhausted the 4 GB of memory we allow per instance.

[9] This is a structural statement about when large graphs remain tractable, not a claim about how often practical queries are localized.

[10] The 600 s limit is passed to the solver, which checks it only at node boundaries. This is why the median of the interception group exceeds 600 s.

Baselines. Once the potencies have been computed, the optimal set is found by exhaustive enumeration over the affordable subsets. One could instead use a cheaper selection rule. On the same 8-experiment instance with $B = 2$, greedy selection (repeatedly adding the experiment with the best potency gain per cost) matched the exhaustive optimum exactly; picking the single most potent experiment that fits the budget reached 53% of it; choosing an affordable subset at random reached 22%.[11] Greedy's strong showing is not a guarantee: potency is not additive across experiments—in Table 1, the best singleton $\tilde{a}_6$ is not part of the optimal budget-3 set $\{\tilde{a}_4, \tilde{a}_5\}$, so settling for the single best experiment reaches only 64% of the optimum there—and experiments that are useless alone can matter in combination.

J.3 Incompleteness of the Pruning Rules

The pruning rules are one-sided certificates: they never discard a useful experiment, but they might keep some useless ones. To quantify this, we evaluated every surviving candidate numerically on the $|\boldsymbol{V}| = 5$ benchmark above (26 instances). In 10 of them, pruning eliminated every candidate, so no program had to be solved at all. The remaining 16 contribute 102 candidates in total, of which 48 survived both rules. Six of these did not finish within the timeout; of the 42 that did, 34 turned out to have zero potency anyway (81%).

K Experimental Setup Details

This appendix details the synthetic evaluation protocol of Section 7.

K.1 Common Parameters

Both the Erdős–Rényi and *bnlearn* evaluations share the following configuration:

$\lvert\mathcal{A}\rvert$ (candidate experiments per simulation)	200
Master seed	42
θ `CF_fraction`	1.0
θ outcome size $\lvert\boldsymbol{Y}^{(\mathfrak{c})}\rvert$	1
θ intervention size $\lvert\boldsymbol{X}^{(\mathfrak{c})}\rvert$	1
Experiment `CF_fraction`	0.3
Experiment outcome size $\lvert\boldsymbol{W}^{(\mathfrak{c})}\rvert$	$\mathrm{Unif}\{1,2,3\}$
Experiment intervention size $\lvert\boldsymbol{Z}^{(\mathfrak{c})}\rvert$	$\mathrm{Unif}\{1,2,3\}$

We fix θ to be a two-world counterfactual to guarantee that the query is not point-identifiable, placing every simulation in the partial-identification regime our framework targets. The 30% counterfactual share for candidate experiments reflects a plausible mix of interventional and counterfactual experiments a practitioner might consider.

K.2 Query Sampling

Each query θ is sampled as follows:

1. Pick $Y^{(1)}$ uniformly from $\boldsymbol{V}$; pick $X^{(1)}$ uniformly from $\mathrm{anc}(Y^{(1)}) \setminus \{Y^{(1)}\}$.
2. For the second world, pick $Y^{(2)}$ uniformly from $\boldsymbol{V}$ subject to $X^{(1)} \in \mathrm{anc}(Y^{(2)})$, then set $X^{(2)} = X^{(1)}$.

K.3 Candidate Experiment Sampling

Each $\tilde{a} \in \mathcal{A}$ is sampled independently. With probability 0.3, $\tilde{a}$ has two counterfactual worlds; otherwise one.

1. Sample $|\boldsymbol{W}^{(1)}|, |\boldsymbol{Z}^{(1)}| \sim \mathrm{Unif}\{1,2,3\}$.

[11] However, the search strategy is not where the cost lives—greedy, random, and exhaustive all rest on the same potency evaluations, and pruning cuts that shared cost for all of them alike.

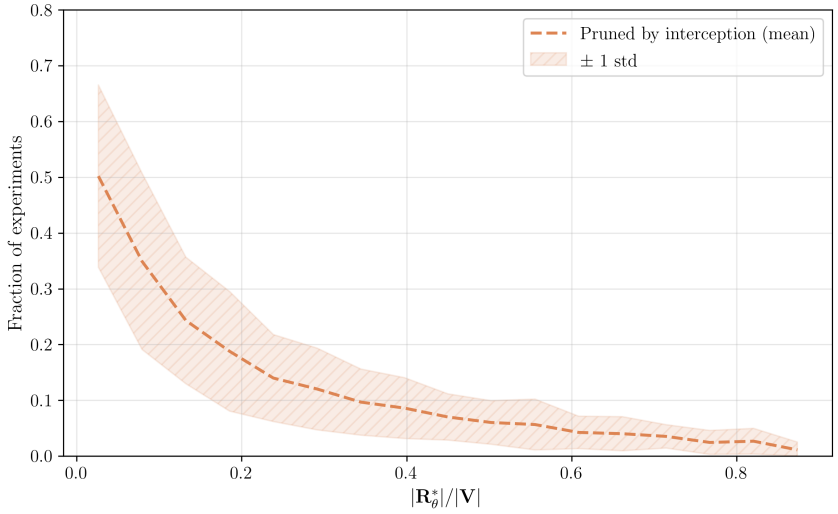


Figure 9: *bnlearn*. The fraction of experiments pruned by the interception mechanism (Algorithm 4) against the relative size of $\boldsymbol{R}^*_\theta$. The line represents the mean fraction of experiments pruned; the band around it represents the ± 1 standard deviation.

2. Pick $\boldsymbol{W}^{(1)} \subseteq \boldsymbol{V}$ uniformly; pick $\boldsymbol{Z}^{(1)} \subseteq \mathrm{anc}(\boldsymbol{W}^{(1)}) \setminus \boldsymbol{W}^{(1)}$ uniformly.
3. For two-world experiments, sample $\boldsymbol{W}^{(2)}, \boldsymbol{Z}^{(2)}$ analogously, but seed each with a shared element: pick $w \in \boldsymbol{W}^{(1)}$ uniformly and require $w \in \boldsymbol{W}^{(2)}$; pick $z \in \boldsymbol{Z}^{(1)} \cap (\mathrm{anc}(\boldsymbol{W}^{(2)}) \setminus \boldsymbol{W}^{(2)})$ uniformly and require $z \in \boldsymbol{Z}^{(2)}$. Remaining variables fill from $\boldsymbol{V}$ and $\mathrm{anc}(\boldsymbol{W}^{(2)}) \setminus \boldsymbol{W}^{(2)}$, respectively.

K.4 Erdős–Rényi Graphs

Each ER graph has a fixed topological order $V_1 < \cdots < V_N$. Directed edges $V_i \to V_j$ for $i < j$ are drawn independently with probability p; bidirected edges between each pair are drawn independently with probability q.

$N = \lvert\boldsymbol{V}\rvert$	$\{10, 15, 20, 30\}$
p	$0.5 \cdot \log(N)/N$
q	c/N, with $c \in \{0.5, 1.0, 1.5, 2.0\}$
Simulations per (N, c)	100

The full sweep yields $4 \times 4 \times 100 = 1{,}600$ simulations, of which 1,596 are reported; in four the sampled graph admitted no query satisfying Appendix K.2, and the simulation was skipped.

K.5 *bnlearn* Graphs

We use 11 networks from the *bnlearn* repository: `ASIA`, `SACHS`, `ALARM`, `BARLEY`, `CHILD`, `INSURANCE`, `MILDEW`, `WATER`, `HAILFINDER`, `HEPAR2`, `WIN95PTS`. Only the graph structure is retained; conditional probability tables are discarded. Since these DAGs are not associated with latent variables, confounders are introduced synthetically. For each simulation, after sampling θ on the bare DAG:

1. Add a forced confounder between $X^{(1)}$ and $Y^{(1)}$.
2. Draw a target ratio $r \sim \mathrm{Unif}(0.1, 0.9)$, clamped below by $1/|\boldsymbol{V}|$ for achievability.
3. Add confounders between uniformly random pairs of observed variables until the total reaches $\lfloor r \cdot |\boldsymbol{V}| \rfloor$.

With 100 simulations per graph, this yields $11 \times 100 = 1{,}100$ simulations.

K.6 Compute Used

All experiments were executed on a single Microsoft Surface Laptop Studio with an 11th Gen Intel Core i7-11370H CPU (3.30 GHz base) and 32 GB RAM. No GPU was used and no parallelization was employed; all simulations ran single-threaded. Each of the three experimental components—the Erdős–Rényi sweep (1,596 simulations), the *bnlearn* sweep (1,100 simulations across 11 networks), and the NHANES end-to-end demonstration (Section 7, Table 1)—completed in a matter of hours. No additional compute infrastructure was used.

K.7 Per-graph results (*bnlearn*)

Table 6 reports the per-graph mean and standard deviation of pruning rates across 100 simulations per graph, broken down by mechanism. Across all 1,100 simulations, total pruning has mean 63.4% (std 12.4%), of which $47.1\% \pm 14.5\%$ is contributed by the ID check and $16.2\% \pm 17.5\%$ by interception. The interception rate varies by more than an order of magnitude across networks—from $2.3\% \pm 4.3\%$ on `MILDEW` to $29.2\% \pm 25.5\%$ on `WIN95PTS`—driven primarily by the relative size of the district hull $|\boldsymbol{R}_\theta^*|/|\boldsymbol{V}|$ (Figure 9).

Table 6: Per-graph pruning rates on *bnlearn* networks (mean $\pm$ std over 100 simulations).

Graph	Interception	ID check	Total
`ASIA`	0.170 ± 0.134	0.332 ± 0.130	0.502 ± 0.158
`SACHS`	0.226 ± 0.194	0.316 ± 0.113	0.543 ± 0.152
`INSURANCE`	0.073 ± 0.073	0.519 ± 0.090	0.592 ± 0.113
`MILDEW`	0.023 ± 0.043	0.574 ± 0.085	0.597 ± 0.092
`WATER`	0.138 ± 0.136	0.490 ± 0.102	0.628 ± 0.092
`BARLEY`	0.122 ± 0.122	0.517 ± 0.098	0.639 ± 0.083
`CHILD`	0.221 ± 0.181	0.428 ± 0.124	0.649 ± 0.113
`ALARM`	0.173 ± 0.173	0.510 ± 0.132	0.682 ± 0.070
`HEPAR2`	0.104 ± 0.116	0.579 ± 0.095	0.682 ± 0.064
`HAILFINDER`	0.243 ± 0.193	0.475 ± 0.137	0.718 ± 0.073
`WIN95PTS`	0.292 ± 0.255	0.446 ± 0.179	0.738 ± 0.085

K.8 Per-density breakdown (Erdős–Rényi)

Table 7 reports the breakdown of pruning rates across the four confounding-density settings $c \in \{0.5, 1.0, 1.5, 2.0\}$, where edges of the bidirected graph are drawn independently with probability c/N (Appendix K.4). The mean achieved confounder ratio $|\boldsymbol{U}|/|\boldsymbol{V}|$ grows monotonically with c, from 0.24 at c=0.5 to 0.94 at c=2.0. As confounding density increases, interception weakens (mean drops from 70.9% to 16.5%) and the ID check picks up the slack (mean rises from 17.2% to 45.6%), but total pruning still declines from 88.1% to 62.1%.

Table 7: Per-density breakdown for the Erdős–Rényi evaluation (mean $\pm$ std). $|\boldsymbol{U}|/|\boldsymbol{V}|$ is the mean confounder ratio for the given parameter c; remaining columns are pruning rates by mechanism.

c	$\lvert\boldsymbol{U}\rvert/\lvert\boldsymbol{V}\rvert$	Interception	ID check	Total
0.5	0.240	0.709 ± 0.310	0.172 ± 0.190	0.881 ± 0.136
1.0	0.472	0.475 ± 0.348	0.304 ± 0.216	0.780 ± 0.163
1.5	0.699	0.300 ± 0.345	0.389 ± 0.218	0.689 ± 0.185
2.0	0.942	0.165 ± 0.300	0.456 ± 0.201	0.621 ± 0.184

K.9 ID-only pruning (subset search)

For the subset search space reduction reported in Section 7, we run the ID check on all $|\mathcal{A}|$ candidates independently, without prior interception filtering. Tables 8 and 9 report the results across the same data-generating regime.

Table 8: ID-only pruning on *bnlearn* networks (mean $\pm$ std over 100 simulations).

Graph	$\lvert \boldsymbol{V} \rvert$	ID-only
ASIA	8	0.466 ± 0.146
SACHS	11	0.479 ± 0.121
CHILD	20	0.612 ± 0.093
INSURANCE	27	0.582 ± 0.093
WATER	32	0.600 ± 0.074
MILDEW	35	0.611 ± 0.088
ALARM	37	0.634 ± 0.064
BARLEY	48	0.618 ± 0.063
HAILFINDER	56	0.660 ± 0.050
HEPAR2	70	0.664 ± 0.050
WIN95PTS	76	0.669 ± 0.046
Overall		0.600 ± 0.108

Table 9: ID-only pruning for Erdős–Rényi, pooled over $\lvert \boldsymbol{V} \rvert$ (mean $\pm$ std).

c	$\lvert \boldsymbol{U} \rvert / \lvert \boldsymbol{V} \rvert$	ID-only
0.5	0.217	0.673 ± 0.060
1.0	0.473	0.637 ± 0.093
1.5	0.719	0.595 ± 0.113
2.0	0.942	0.552 ± 0.134
Overall		0.614 ± 0.113

L NHANES Experiment Details

This appendix details the construction of the NHANES experiment in Section 7. We use the NHANES 2017–2018 cycle[12] and restrict the sample to adult respondents (RIDAGEYR ≥ 18) with non-missing entries on Z, X and Y, yielding $n = 5{,}833$ complete cases.

The assumed causal structure (Figure 4 in Section 7) encodes the following substantive assumptions: balance problems (A) may cause falls (B), and both may influence the propensity to engage in physical activity (X); health insurance (Z) acts as an instrument-like predictor of physical activity through access to preventive care; and physical activity affects the (negated) diabetes outcome (Y). Three latent confounders are posited: U_3 between A and B (e.g., inner-ear or vestibular disorders affecting both balance and falls), U_1 between Z and X (e.g., overall healthcare access and socioeconomic factors), and U_2 between X and Y (e.g., genetic predisposition or general health-consciousness). The query of interest is the average treatment effect $\theta = \mathrm{ATE} = P(y_x) - P(y_{x'})$ of physical activity on not having diabetes.

Table 10 maps each variable to its corresponding NHANES variable code and the rule used to discretize it to a binary value. We negate the diabetes indicator (denoting the variable $\neg Y$ in the table) so that $Y = 1$ corresponds to *not* having a doctor-diagnosed diabetes condition; this preserves the natural reading of $\theta > 0$ as "physical activity is protective."

[12] https://wwwn.cdc.gov/nchs/nhanes/

Table 10: Mapping of variables to NHANES 2017–2018 codes. A and B are part of the assumed structure but are not recorded in this cycle.

Node	Variable	Code	Discretization rule	$\lvert\text{supp}\rvert$
A	Balance problems	—	not recorded in this cycle	2
B	Falls	—	not recorded in this cycle	2
Z	Health insurance	HIQ011	1 if currently covered	2
X	Physical activity	PAQ650/665	1 if vigorous or moderate recreation	2
$\neg Y$	Diabetes	DIQ010	1 if doctor-diagnosed	2

The five-variable ADMG is the input to the pruning step, which uses only the graph and the query. The potency programs take the discretized observational distribution $P(Z, X, Y)$.[13]

L.1 Candidate Experiments: Realizable Trials vs. Assumption-Type Constraints

As defined in Section 3, the candidate set $\mathcal{A}$ includes both physically realizable experiments and assumption-type constraints; the polynomial program treats both uniformly, and the distinction is captured by the cost function, which may assign infinite cost to experiments that cannot be conducted (Section 4). Whether a given counterfactual quantity is physically realizable can be checked upfront with the algorithm of Raghavan and Bareinboim [2025]. Table 11 labels each candidate of Table 1 accordingly.

Table 11: Candidates of Table 1, labeled as realizable trials or assumption-type constraints.

	Experiment	Type	Interpretation
$\tilde{a}_1$	$P(X \mid do(B))$	realizable	single-world intervention
$\tilde{a}_2$	$P(B \mid do(A))$	realizable	single-world intervention
$\tilde{a}_3$	$P(X, Y \mid do(z))$	realizable	single-world intervention
$\tilde{a}_4$	$P(y_{x'}, y_x')$	assumption	cross-world counterfactual
$\tilde{a}_5$	$P(Y \mid do(x'))$	realizable	single-world intervention
$\tilde{a}_6$	$P(Y \mid do(x))$	realizable	single-world intervention

Concretely, $\tilde{a}_4 = P(y_{x'}, y'_x)$ is the probability that physical activity (necessarily and sufficiently) *causes* diabetes. No trial can measure it, but committing to $\tilde{a}_4 = 0$ on substantive grounds is exactly the monotonicity ("no harm") assumption, which is known to tighten PNS bounds [Tian and Pearl, 2000]. Its cost of 1 in Table 1 prices that commitment, not a trial. The combinations $\{\tilde{a}_4, \tilde{a}_6\}$ and $\{\tilde{a}_4, \tilde{a}_5\}$ mix both types.

L.2 Computational Procedure

Every potency in Table 1 is the optimum of the single coupled polynomial program of Algorithm 3 (Appendix G.2); the experimental outcome is never discretized. Programs are constructed following the procedure of Duarte et al. [2024], using the implementation in the supplementary material, and solved by Couenne at its default tolerances—no custom optimality gap, and a 300-second time limit that no solve came close to—i.e. to global optimality up to solver tolerance; early termination plays no role in any reported number. The computation is reproducible from `nhanes_pipeline.ipynb` in the supplementary material.

L.3 Robustness

Finite-sample stability. The pipeline uses the plug-in principle and treats the empirical distribution of the $n = 5{,}833$ NHANES records as exact. Methodologically, sampling error can be accounted for within the same machinery: Duarte et al. [2024, Section 6] relax each observational equality constraint to a $1-\alpha$ confidence region, and since our experimental constraints and the potency program are built

[13] A and B are not recorded in this cycle, so they do not appear in $P(\boldsymbol{V})$. Since they influence the remaining variables only through X and share no latent parent with Z or Y, marginalizing them out leaves exactly the ADMG over (Z, X, Y) used here, so these are the bounds of the five-variable model.

on the same constraint set, the relaxation carries over unchanged; the pruning step uses only the graph and the query, never $P(\boldsymbol{V})$, so pruning decisions are unaffected by sampling error by construction. To quantify what the plug-in treatment neglects, we bootstrapped the sample (50 replicates) and recomputed the potencies. The observational width averaged 0.86 ± 0.03 across replicates (range 0.82 to 0.94), and all potencies had standard deviations below 0.025 (e.g., $\tilde{a}_3$: 0.000 ± 0.000; $\tilde{a}_4$: 0.438 ± 0.009; $\{\tilde{a}_4, \tilde{a}_5\}$: 0.774 ± 0.024). The optimal choice at $B = 1, 2, 3$ agreed with Table 1 in all 50 replicates. The potency programs use only $P(Z, X, Y)$, whose 8 cells are estimated from $5{,}833$ records, so this stability is expected.

Graph misspecification. Dependence on the assumed graph is fundamental to any method in this space, but our framework separates two kinds of dependence. Pruning depends on the graph only through the district hull $\boldsymbol{R}^*_\theta$ of the query. The potency values depend on more than that: by Lemma 6.10 they are determined by the hull together with the mechanisms of the variables inside it, so an edit that changes one of those mechanisms can change the potencies even when the hull itself is untouched.

Adding a direct edge $Z \to Y$—dropping the assumption that health insurance affects diabetes only through physical activity—separates the two. The hull $\boldsymbol{R}^*_\theta$ is unchanged, and so is every pruning decision in Table 1: $\tilde{a}_1$ is still certified by the ID check and $\tilde{a}_2$ still by the interception rule. But the edge makes Z a parent of Y, and Y lies inside the hull, so the existing potencies change. The observational width rises from 0.86 to 1.00, because the bounds can no longer use the assumption that Z acts on Y only through X, and every potency rises with it: $\tilde{a}_4$ from 0.44 to 0.48, $\tilde{a}_6$ from 0.50 to 0.53, and $\{\tilde{a}_4, \tilde{a}_5\}$ from 0.78 to 0.85. The optimal choice at each budget $B = 1, 2, 3$ is nevertheless the same under both graphs.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction summarize the max-potency formulation, NP-hardness, potency evaluation, pruning criteria, and empirical results developed in Sections 4–7. The scope is restricted to discrete variables.

   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: Section 8 discusses the limitations explicitly.

   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: Proofs are provided in Appendices E and H, with auxiliary lemmas in Appendix I. Assumptions are stated in Section 2 and alongside the results.

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: Appendix K describes the synthetic evaluation protocol, including master seed 42. Appendix L specifies NHANES preprocessing and potency computation. The supplement provides code, notebooks, and data extracts.

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: The supplement provides the implementation, experiment notebooks, benchmark graphs, and NHANES data and preprocessing code. Environment details are given in `README.md` and `requirements.txt`.

   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: Synthetic settings are specified in Appendix K. NHANES preprocessing, solver settings, and robustness checks are given in Appendix L. No models are trained.

   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: Section 7 and Appendix K report mean pruning rates with ± 1 sample standard deviation across simulations. Appendix L reports a 50-replicate NHANES bootstrap, including potency variability and selection stability. These standard deviations describe variability, not confidence intervals.

   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: Appendix K.6 reports the hardware, single-threaded execution setup, and approximate runtimes.

   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: We have reviewed the NeurIPS Code of Ethics. The work uses synthetic graphs and public datasets; no participants were recruited or new human-subject data collected.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: The method supports transparent decisions by quantifying causal uncertainty. Incorrect graph assumptions or treating empirical distributions as exact can lead to overstated guarantees; these limitations are discussed in Section 8 and Appendix L.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The released code and notebooks use public benchmark structures and NHANES data. No pretrained models or other high-risk assets are released.

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: We credit the *bnlearn* repository (`https://www.bnlearn.com/bnrepository/`) and NHANES 2017–2018 public-use data (`https://wwwn.cdc.gov/nchs/nhanes/`). The supplement includes adapted `Ananke` components (Apache License 2.0, `https://gitlab.com/causal/ananke`) for the ID algorithm of Shpitser and Pearl [2006], and `autobound` (MIT License) from Duarte et al. [2024], with original licenses retained.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The supplement documents the code, notebooks, and data extracts in `README.md`, with dependencies in `requirements.txt`. Our code is released under the MIT License; third-party licenses are retained.

    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: No crowdsourcing or participant recruitment was conducted. We use existing public NHANES data.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: The study uses existing public, de-identified NHANES data and involves no new participant recruitment or data collection.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not a component of the core research methods.

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.