COMPUTATIONAL IDENTIFIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Identification conditions describe the computability of a target query or parameter of interest as a function of the type and amount of information available. In causal identification, this information is often expressed in the form of a causal graph, and data are observed or collected for some subset of variables in the graph. Target queries may be for a single effect alone or for a class of effects in a given model. The derivation of an identification algorithm then defines mathematically the process by which the desired causal effect(s) can be uniquely determined, theoretically, in expectation. Identifiability in expectation, or 'theoretical identifiability,' generally assumes asymptotic properties, infinite data, or other mathematically idealized conditions. In this paper, we explore a fundamental distinction between this theoretical, idealized notion of identifiability and a proposed alternative that is computation-bound. The framework we propose — 'computational identifiability' — is to instead define a finite computational search procedure for an empirical estimator. If this process finds an estimator empirically within a desired error tolerance, then identifiability is satisfied, *conditional on the specified assumptions of the search (i.e., a prior distribution over the parameters) and conditional on the search procedure itself.* Through several experiments, we demonstrate how this framework allows us to answer fine-grained, practical identification questions that are difficult to analyze theoretically, such as identification with small finite samples, with ambiguous graphical criteria, with mixed observational-interventional data, and across counterfactual data and estimands.

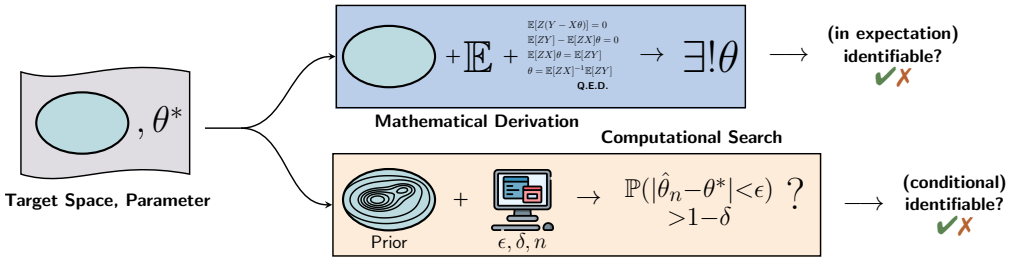


Figure 1: Two approaches to identifiability. Mathematical derivation (top) seeks to prove the existence of a unique parameter, analytically in expectation. Computational search (bottom, proposed in this work) instead defines an empirical search procedure for an estimator and defines 'computational identifiability' as the successful discovery of an estimator, conditional on a prior over parameters, a hypothesis space over estimators, a finite sample size, and a desired error tolerance.

1 INTRODUCTION

A critical step in being able to answer a causal or statistical query from data is to distinguish whether or not there is sufficient information to determine a unique answer. Identifiability addresses this problem, deriving mathematically what can be determined from a collection of distributions (and assumptions), in expectation. In causal inference, identification algorithms make use of causal graphical criteria, probability axioms, matrix or algebraic equations, bounding arguments, and other mathematical tools. This notion of identifiability in expectation, which we term 'theoretical identifiability,' generally assumes asymptotic properties, infinite data, or other mathematically idealized conditions.

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

While theoretical identifiability is a long-solved problem for many model classes in expectation, with sound and complete algorithms like do-calculus (Pearl & Robins, 1995; Shpitser & Pearl, 2008), po-calculus (Malinsky et al., 2019), or σ-calculus (Correa & Bareinboim, 2020a), identifiability itself is far from a solved problem. Identification in new settings requires bespoke analysis (Puli & Ranganath, 2020; Puli et al., 2020). Moreover, many important settings involve finite sample sizes, ambiguous graphical criteria, or complex combinations of observational and experimental data that defy the requirements of existing algorithms and known bounds. In such scenarios, theoretical guarantees of identifiability in expectation may offer little guidance on whether a target parameter can actually be estimated. Similarly, many settings considered unidentifiable nonparametrically may become identifiable under parametric assumptions (such as linearity, non-Gaussianity, or monotonicity (Shimizu et al., 2006; Puli & Ranganath, 2020)). Exploring these possibilities analytically, especially for nonlinear settings, remains a significant challenge. Novel settings that do not yet have theoretical results, or where identifiability assumptions are hard to satisfy, leave the search for new identification algorithms and partial identification bounds open and ongoing (Raghavan & Bareinboim, 2026; Ghassami et al., 2023; Tan et al., 2024; Kline & Tamer, 2023; Lee & Bareinboim, 2020).

In this work, we introduce an alternative notion of identifiability: 'computational identifiability,' that instead frames identifiability as a finite computational search (see Figure 1). In lieu of mathematical derivation, computational identifiability starts with (1) a meta-prior over the parameters in question: in this case, a distribution over structural causal models (SCMs); and (2) a hypothesis space of possible estimators. For a given finite sample size n, desired error tolerance ϵ, and desired confidence bound δ, computational identifiability defines successful identification as the existence of a satisfactory estimator in the given hypothesis space. In other words, computational identifiability defines a practical and empirical, hence actionable, notion of identifiability.

Contributions. Our main contributions are the following.

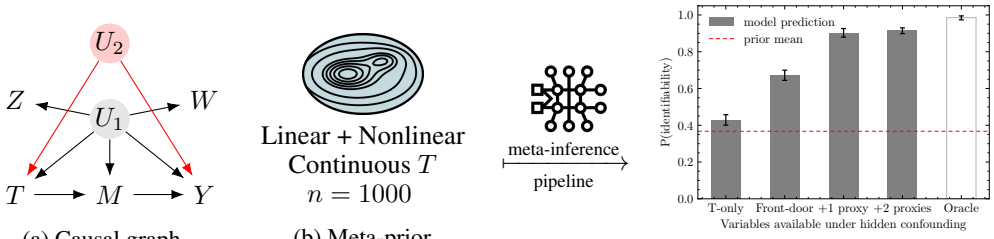
1. We introduce **computational identifiability**, a practical, computation-bound notion of identifiability to complement existing theoretical and idealized notions.
2. We formalize and further generalize the connection between causal effect estimation and meta-learning (Section 3), defining the required components to compute and evaluate computational identifiability (hypothesis space $\mathcal{H}$, error tolerance ϵ, confidence bound δ, and mixture distributions for custom meta-priors over SCMs and empirical data).
3. We empirically demonstrate computational identifiability across a variety of complex and small finite sample settings, including proximal front-door identification (Section 2), identification with ambiguous graphical criteria (Section 4.1), with mixed observational-interventional data (Section 4.2), and across counterfactual data and estimands (Section 4.3).

2 Computational identification in practice: proximal front-door 'metadentifiability'

We first illustrate a concrete example of how *computation (via meta-inference) can allow us to derive a new type of identifiability result*. This example is inspired by recently published identifiability results in Guo et al. (2026) that tackle an important causal inference setting for which identification remains an actively researched problem: confounded front-door causal structures (e.g., Figure 2a).

This causal structure arises frequently in practice, for example in medical settings, where mediators M often share hidden common causes with treatments T and outcomes Y, violating the assumptions required for identification via the front-door criterion (which would otherwise allow for identification under hidden confounding). The idea of Guo et al. (2026) and others (e.g., Dukes et al., 2023; Bai et al., 2025) is to develop proximal generalizations of the front-door criterion for such settings, building on proximal identification theory (Shpitser et al., 2023; Kuroki & Pearl, 2014; Miao et al., 2018) to achieve identification using proxies of U_1 (here, Z and W) in addition to mediator M.

These works are great examples of the *theoretical approach* to identification, where we work analytically to come up with possible sets of assumptions under which the desired treatment effect (or interventional distribution) can be successfully written as a functional of the observational distribution, and we then use those derived assumptions to guide the analytical development of corresponding estimation strategies. Specifically for Figure 2a, Guo et al. (2026) have to derive rank conditions



(a) Causal graph (b) Meta-prior (c) Empirical identifiability

Figure 2: Computational identification can empirically lower-bound a 'probability of identifiability' in cases where theoretical results may be difficult, inapplicable, or unknown: in this case, proximal generalization of the front-door criterion (including single-proxy control) with continuous treatment.

on and completeness conditions for conditional expectations of the relevant variables (and square-integrable functions thereof), relying also on solutions to Fredholm integral equations of the first kind and/or boundedness and distinctness conditions. They prove identification asymptotically (and then separately develop estimation approaches) for cases where these sets of assumptions are met.

Importantly, existing theory in Guo et al. (2026) for this setting assumes binary treatment and access to both proxies. But what about cases with continuous treatment, which might be common in a medical setting? What if we can only measure one proxy? Is a continuous treatment effect identifiable?

The pipeline shown in Figure 2 typifies the alternate, complementary approach to identification that we develop in this paper to help address such questions on the edges of our known theories: a *computational approach*. Computational identification starts with a meta-prior that describes the settings we want to consider. In Figure 2b, we define this meta-prior as a distribution over Figure 2a with random linear $\{T, M\}$, random nonlinear $\{Z, W, Y\}$, and Gaussian $\{U_1, U_2\}$ (full details in Appendix E.1). We then *use meta-inference to automatically search over a hypothesis space $\mathcal{H}$ of dataset-level estimators*. In Figure 2, $\mathcal{H}$ is defined by the expressivity of a transformer-based neural network that we meta-train. The meta-model then learns (or fails to learn) to estimate the average treatment effect, given only observed datasets of size n as input. This pipeline effectively translates our theoretical identification question into a computational question: the meta-model's performance encodes a *lower-bound to the true analytical identifiability* in the input setting.

In this manner, *computational identifiability has allowed us to swap the mathematical derivation of identification conditions (e.g., rank conditions, completeness conditions, Fredholm integral equations) with empirical hyperparameter optimization*. Importantly, modeling assumptions in both the theoretical and computational case are untestable; as in all causal inference settings, the goal is not to completely eliminate untestable assumptions, but instead, to trade one set of untestable assumptions for a different set that we might find more suitable (or easier to use) for our given application. Computational identifiability allows us to, for example, derive *finite-sample identification results* in any setting we can describe via simulation. In general, settings that are particularly challenging to solve analytically may lend themselves more easily to computational exploration.

Figure 2c shows five such identifiability results, corresponding to five different sets of observed variables: t-only ($\{T, Y\}$), front-door ($\{T, M, Y\}$), one proxy ($\{W, T, M, Y\}$), two proxies ($\{Z, W, T, M, Y\}$), and the oracle case ($\{U_1, U_2, T, M, Y\}$), all with continuous treatments. Loosely speaking, each value in Figure 2c answers the question "What is the probability that identifiability is satisfied under these assumptions?" — an empirical answer to our otherwise theoretical question (we formalize this probabilistic notion in Section 3). An interesting takeaway in the case of Figure 2c: the causal effect is about as identifiable with one proxy as it is with two.

Despite important theoretical differences between these settings (e.g., single-proxy, two-proxy, and continuous proximal identification each have their own streams of literature (Tchetgen et al., 2020; Bozkurt et al., 2026; Vollmer et al., 2026; Xu & Gretton, 2025; Park et al., 2023; Wu et al., 2024)), *computational identifiability allows us to explore each of them using the same approach*.

In the remainder of the paper, we introduce the formalisms that generalize this approach and allow us to implement computational identification across a variety of important causal inference settings.

3 COMPUTATIONAL IDENTIFIABILITY

Notation. We provide a full set of preliminaries in Appendix A and briefly introduce notation here. A structural causal model (SCM) $\mathfrak{C} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$ is defined as a 4-tuple with observed variables $\mathbf{V}$, unobserved variables $\mathbf{U}$, functions $f_i \in \mathbf{F}$ for each $V_i \in \mathbf{V}$, and a distribution $P_{\mathbf{U}}$ over $\mathbf{U}$. SCM $\mathfrak{C}$ entails *observational distribution* $P^{\mathfrak{C}}$. We denote $\mathfrak{C}$ after intervention I as $\mathfrak{C}^{\text{do}(I)}$, and the resulting *interventional distribution* as $P^{\mathfrak{C};\text{do}(I)}$. We denote the *counterfactual distribution* $P^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\text{do}(I)}$ after intervention I given factual evidence (i.e., already-observed data) $\mathbf{Z} = \mathbf{z}$ where $\mathbf{Z} \subseteq \mathbf{V}$.

How do we describe causal identification across settings as a computational process? We first define several objects that allow us to precisely describe joint mixtures of queries and data that may simultaneously be empirical and/or SCM-based observational, interventional, and counterfactual.

Definition 1 (Causal query)**.** *Let $\mathbb{C}$ denote the space of all SCMs with a fixed set of endogenous variables $\mathbf{V}$. We define a* causal query $\Theta[\cdot]$ *as a functional taking as input an SCM $\mathfrak{C} \in \mathbb{C}$ as well as a query point $\mathbf{X} = \mathbf{x}$ for any $\mathbf{X} \subseteq \mathbf{V}$. More formally, let $\Omega_{\mathbf{X}}$ denote the sample space for any subset $\mathbf{X} \subseteq \mathbf{V}$ and define $\Omega = \bigsqcup_{\mathbf{X} \subseteq \mathbf{V}} \Omega_{\mathbf{X}}$. Then $\Theta : \mathbb{C} \times \Omega \to \mathbb{R}$.*

For example, average treatment effects such as the population average treatment effect (PATE) and conditional average treatment effect (CATE) can be defined as expectations of interventional distributions specific to a given SCM $\mathfrak{C}$ that either make use of or ignore a given observation $\mathbf{x}$. Specifically, given SCM $\mathfrak{C} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$ with a binary treatment variable $T \in \mathbf{V}$ and outcome variable $Y \in \mathbf{V}$, the PATE can be written as $\Theta_{\text{PATE}}[\mathfrak{C}, \cdot] := \mathbb{E}_{p_Y^{\mathfrak{C};\text{do}(T=1)}}[Y] - \mathbb{E}_{p_Y^{\mathfrak{C};\text{do}(T=0)}}[Y]$, where the second argument is ignored. Adding a condition $\mathbf{X} = \mathbf{x}$ for conditioning variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y, T\}$, the CATE can be written as $\Theta_{\text{CATE}}[\mathfrak{C}, \mathbf{x}] := \mathbb{E}_{p_{Y|\mathbf{X}=\mathbf{x}}^{\mathfrak{C};\text{do}(T=1)}}[Y] - \mathbb{E}_{p_{Y|\mathbf{X}=\mathbf{x}}^{\mathfrak{C};\text{do}(T=0)}}[Y]$. Common estimands comprised of counterfactual rather than interventional distributions include individual treatment effects (ITEs) and sample average treatment effects (SATEs). The ITE represents the counterfactual difference in outcomes specific to a given unit with observed features $\mathbf{V} = \mathbf{v}$ (including realized T and Y values): $\Theta_{\text{ITE}}[\mathfrak{C}, \mathbf{v}] := \mathbb{E}_{p_Y^{\mathfrak{C}|\mathbf{V}=\mathbf{v};\text{do}(T=1)}}[Y] - \mathbb{E}_{p_Y^{\mathfrak{C}|\mathbf{V}=\mathbf{v};\text{do}(T=0)}}[Y]$. Building on the ITE, the SATE represents the average ITE across a given set of observed units, which can be viewed as a composition of a set of causal queries, i.e., given dataset $\mathcal{D} = \{\mathbf{v}_i\}_{i=1}^n$, the SATE is $\frac{1}{n}\sum_{i=1}^n \Theta_{\text{ITE}}[\mathfrak{C}, \mathbf{v}_i]$.

Definition 2 (Causal mixture distribution)**.** *Let $\mathbb{I} = \{\emptyset, I_1, I_2, ..., I_K\}$ represent a set of interventions indexed by k, where $\emptyset$ represents the null-intervention (doing nothing). Let $\mathbb{I}_{\mathbb{D}} \subseteq \mathbb{I}$ represent the set of interventions for which empirical data are collected. For example, if empirical data are all observational, $\mathbb{I}_{\mathbb{D}} = \{\emptyset\}$. If in addition to observational data, data are collected from a randomized control trial conducted with a treatment intervention $I_{treated}$, a control intervention $I_{control}$, and a placebo intervention $I_{placebo}$, then $\mathbb{I}_{\mathbb{D}} = \{\emptyset, I_{treated}, I_{control}, I_{placebo}\}$. Let $\mathbb{D} = \{\mathcal{D}_j\}_{j=1}^{|\mathbb{I}_{\mathbb{D}}|}$ represent the corresponding set of empirical datasets (multisets). We also index $\mathbb{I}_{\mathbb{D}}$ with j such that $\mathcal{D}_j$ corresponds to intervention $I_j \in \mathbb{I}_{\mathbb{D}}$. The observed variables (observed in each dataset $\mathcal{D}_j \in \mathbb{D}$) are denoted $\mathbf{O}$. We then define a* causal mixture distribution *given SCM $\mathfrak{C}$ over endogenous variables $\mathbf{V}$ where $\mathbf{O} \subseteq \mathbf{V}$ as a weighted sum of empirical distributions and SCM-based observational, interventional, and counterfactual distributions:*

$$p^{\mathfrak{C}}_{\mathbf{O},mix}(\mathbf{o}) := \sum_{I_j \in \mathbb{I}_{\mathbb{D}}} \sum_{\mathbf{o}_i \in \mathcal{D}_j} \left(\underbrace{\alpha_{ij}^{emp} \delta(\mathbf{o} - \mathbf{o}_i)}_{\substack{\textit{item-weighted empirical} \\ \textit{distribution collected} \\ \textit{under intervention } I_j}} + \sum_{I_k \in \mathbb{I}} \alpha_{ijk}^{cf} \underbrace{p_{\mathbf{O}}^{\mathfrak{C}^{do(I_j)}|\mathbf{O}=\mathbf{o}_i;do(I_k)}(\mathbf{o})}_{\substack{\textit{counterfactual distribution under} \\ \textit{SCM } \mathfrak{C}^{do(I_j)} \textit{ given evidence } \mathbf{o}_i \\ \textit{and hypothetical intervention } I_k}} \right)$$
$$+ \sum_{I_k \in \mathbb{I}} \alpha_k^{int} \underbrace{p_{\mathbf{O}}^{\mathfrak{C};do(I_k)}(\mathbf{o})}_{\substack{\textit{interventional} \\ \textit{distribution for} \\ I_k \textit{ under SCM } \mathfrak{C}}}$$

with $\sum_{ij} \alpha_{ij}^{emp} + \sum_{ijk} \alpha_{ijk}^{cf} + \sum_k \alpha_k^{int} = 1$, $\alpha_{ij}^{emp} \geq 0$, $\alpha_{ijk}^{cf} \geq 0$, $\alpha_k^{int} \geq 0$, and Dirac delta δ.[1]

Traditional 'theoretical' definitions of causal identifiability are about being able to uniquely determine a causal query from a given (typically observational) distribution or set of distributions.

[1]For valid densities, we consider $p^{\mathfrak{C}}_{\mathbf{O},mix}$ with respect to a base measure that dominates both the Lebesgue measure and the counting measure of the empirical data.

Definition 3 (Theoretical identifiability)**.** *Given a space $\Omega_{\mathfrak{C}}$ over SCMs $\mathfrak{C}$ with endogenous variables $\mathbf{V}$ and a subset of observed variables $\mathbf{O} \subseteq \mathbf{V}$, we say a causal query $\Theta[\mathfrak{C}, \mathbf{x}]$ with query point $\mathbf{X} = \mathbf{x}$ for $\mathbf{X} \subseteq \mathbf{V}$ is* theoretically identifiable *from $\mathbf{O}$ if for any two SCMs $\mathfrak{C}_1, \mathfrak{C}_2 \in \Omega_{\mathfrak{C}}$, $\Theta[\mathfrak{C}_1, \mathbf{x}] = \Theta[\mathfrak{C}_2, \mathbf{x}]$ whenever $P^{\mathfrak{C}_1}_{\mathbf{O}} = P^{\mathfrak{C}_2}_{\mathbf{O}}$.*[2] *A consequence is that Θ can be written as a functional of $P^{\mathfrak{C}}_{\mathbf{O}}$.*

Given our setup, we can instead define a new notion of identifiability that is conditional on a particular prior π over SCMs and a hypothesis space $\mathcal{H}$ over functions derived from a causal mixture distribution.

Definition 4 (Computational identifiability)**.** *Given a prior distribution over SCMs $\mathfrak{C} \sim \pi$, a causal mixture distribution $p^{\mathfrak{C}}_{\mathbf{O},mix}$ over observed variables $\mathbf{O}$ with sample space $\Omega_{\mathbf{O}}$, and causal query $\Theta[\mathfrak{C}, \mathbf{x}]$ with query point $\mathbf{X} = \mathbf{x}$ ($\mathbf{X} \subseteq \mathbf{V}$) with sample space $\Omega_{\mathbf{X}}$, we say Θ is ϵ-δ-*identifiable conditional on prior π and hypothesis space $\mathcal{H}$ *if there exists a function $\phi \in \mathcal{H}$ where $\phi : \Omega_{\mathbf{O}} \times \Omega_{\mathbf{X}} \mapsto \mathbb{R}$ such that $\mathbb{E}_{\mathfrak{C}\sim\pi}\left[\mathbb{E}_{\mathbf{O}\sim p^{\mathfrak{C}}_{\mathbf{O},mix}}\left[\mathbf{1}\left(|\Theta[\mathfrak{C}, \mathbf{x}] - \phi(\mathbf{O}, \mathbf{x})| \leq \epsilon\right)\right]\right] \geq 1 - \delta$. In other words, the query can be approximated within an ϵ margin of error by ϕ with probability at least $1 - \delta$.*

This notion is similar to, e.g., PAC learnability (Valiant, 2026), but with an important distinction: a type difference in the computation of error that leads to the possibility of causal identifiability showing up as irreducible error (information loss) even with a perfect learning algorithm. Information loss is represented via the query $\Theta[\mathfrak{C}, \mathbf{x}]$ taking as input the entire SCM $\mathfrak{C}$, while the model $\phi(\mathbf{O}, \mathbf{x})$ instead takes as input only the subset $\mathbf{O}$ of observed context variables following mixture $p^{\mathfrak{C}}_{\mathbf{O},mix}$.

Although causal query $\Theta[\mathfrak{C}, \mathbf{x}]$ is evaluated for SCM $\mathfrak{C}$ pointwise at query point $\mathbf{X} = \mathbf{x}$, in practice, our computational search for a causal estimator will involve a distribution not only over observed variables $\mathbf{O}$ but also over query points $\mathbf{X}$. We formalize this process via the causal query distribution.

Definition 5 (Causal query distribution)**.** *Given SCM $\mathfrak{C}$ over endogenous variables $\mathbf{V}$, observation $\mathbf{O} = \mathbf{o}$ where $\mathbf{O} \subseteq \mathbf{V}$, and causal query $\Theta[\mathfrak{C}, \mathbf{x}]$ taking as input $\mathfrak{C}$ and query point $\mathbf{X} = \mathbf{x}$ where $\mathbf{X} \subseteq \mathbf{V}$, we define $p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o})$ as a distribution over $\mathbf{X}$ conditional on $\mathfrak{C}$ and $\mathbf{o}$.*

Definition 5 covers several cases. For a fixed query point $\mathbf{X} = \mathbf{x}'$, we would have $p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o}) = \delta(\mathbf{x} - \mathbf{x}')$. For query points sampled from the marginal distribution of $\mathbf{X}$ entailed by $\mathfrak{C}$, we would have $p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o}) = p^{\mathfrak{C}}_{\mathbf{X}}$. For an empirical query like the ITE for observation $\mathbf{O} = \mathbf{o}$, we might instead have $p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o}) = \delta(\mathbf{x} - \mathbf{o})$ (if, e.g., $\mathbf{O}$ and $\mathbf{X}$ represent the same set of variables).

Recent meta-learning approaches to causal effect estimation allow us to amortize the learning of posteriors over causal estimators, based on a specified input prior over SCMs (Bynum et al., 2025; Robertson et al.; Balazadeh et al.). Therefore, an analogous meta-learning approach would allow us to learn a *posterior over such functionals* $\phi(\mathbf{O}, \mathbf{x})$ and in turn estimate the *probability of identifiability*:

$$\text{P(identifiability)} = \mathbb{E}_{\mathfrak{C}\sim\pi}\left[\mathbb{E}_{\mathbf{O}\sim p^{\mathfrak{C}}_{\mathbf{O},mix}}\left[\mathbb{E}_{\mathbf{X}\sim p_{query}}\left[\mathbf{1}\left(|\Theta[\mathfrak{C}, \mathbf{X}] - \phi(\mathbf{O}, \mathbf{X})| \leq \epsilon\right)\right]\right]\right]. \quad (1)$$

Intuitively, computational identifiability (i.e., estimating Equation (1) empirically) combines a few distinct additions to and/or relaxations of theoretical identifiability:

- **Prior-conditioning:** is the distribution of the target parameter, conditional on the information we have, sufficiently concentrated around a unique value?
- **Hypothesis-conditioning/learnability:** does an estimator exist within our chosen hypothesis space that could predict the target parameter, and can we learn such an estimator empirically in practice?
- **Approximation/partial identification:** rather than perfect prediction of the target parameter, can we predict the target parameter within an acceptable error tolerance?

3.1 Meta-learning an estimator ϕ

In order to find an estimator (or see if it is possible to find an estimator), we have to learn a mapping from observations and query points $(\mathbf{O} = \mathbf{o}, \mathbf{X} = \mathbf{x})$ to query values $\Theta = \theta$. After choosing a hypothesis space, one way to find such an estimator is to meta-learn the desired mapping in a supervised manner from many examples (Bynum et al., 2025). To do this in practice, we first define a joint distribution over mixture samples and query values, from which we can stream training data.

[2] We have written this definition using observational distribution $P^{\mathfrak{C}}_{\mathbf{O}}$ by default, but for a different class of distribution, the analogous equality would be between whichever class of distributions is available.

Definition 6 (Joint causal-query-mixture distribution)**.** *Let π represent a distribution over SCMs $\mathfrak{C} \sim \pi$ where all SCMs share the same fixed set of endogenous variables $\mathbf{V}$, and let Θ represent a causal query of interest with query point $\mathbf{X} = \mathbf{x}$ where $\mathbf{X} \subseteq \mathbf{V}$. Assume causal mixture distribution $p^{\mathfrak{C}}_{\mathbf{O},mix}(\mathbf{o})$ and causal query distribution $p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o})$ are specified for each SCM $\mathfrak{C}$. We can then define a* joint causal-query-mixture distribution *over observations $\mathbf{O}$ and query values Θ as*

$$p_{joint}(\mathbf{o}, \mathbf{x}, \theta) = \mathbb{E}_{\mathfrak{C}\sim\pi}\left[p^{\mathfrak{C}}_{\mathbf{O},mix}(\mathbf{o}) \cdot p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o}) \cdot \delta\left(\theta - \Theta\left[\mathfrak{C}, \mathbf{x}\right]\right)\right]$$

allowing for joint samples of datapoints, query points, and causal query values $\{(\mathbf{o}, \mathbf{x}, \theta)\}$.

3.1.1 Architectures suitable for meta-learning of ϕ

To meta-learn ϕ, we consider a couple variants of Conditional Neural Process (CNP) style architectures (Garnelo et al., 2018a;b). Neural Processes (NPs) are a family of flexible meta-learning models that operate on set-valued inputs and produce predictions, with uncertainty estimates, over outputs at arbitrary locations (Ashman et al., 2025). In this section, we briefly describe NPs and map them to our setting as a representative model class that we will use for our analysis in this paper. However, the component of our pipeline that performs meta-learning could alternatively leverage other types of models, such as ST-based architectures (Bynum et al., 2025; Zhang et al., 2022), or tabular foundation model (TFM) style architectures (Qu et al., 2026; Balazadeh et al.; Robertson et al.; Ma et al., 2025).

Ashman et al. (2025, Appx. A) provide a unifying construction of the many variants of CNPs as a sequence of three components. Given context set $\mathcal{D}_c = \{\mathbf{x}_c, \mathbf{y}_c\}$ and target set $\mathcal{D}_t = \{\mathbf{x}_t, \mathbf{y}_t\}$, *encoder* $e : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ encodes each $(\mathbf{x}_{c,n}, \mathbf{y}_{c,n}) \in \mathcal{D}_c$ into a latent representation $\mathbf{z}_{c,n} \in \mathcal{Z}$, *processor* $\rho : (\bigcup_{n=0}^{\infty} \mathcal{Z}^n) \times \mathcal{X} \to \mathcal{Z}$ then processes the embedded context $e(\mathcal{D}_c)$ and target input $\mathbf{x}_t$ to obtain target-dependent output $\mathbf{z}_t \in \mathcal{Z}$, and *decoder* $d : \mathcal{Z} \to \mathcal{P}_{\mathcal{Y}}$ maps the target input to the predictive distribution over the output at that location (where $\mathcal{P}_{\mathcal{Y}}$ is the set of distributions over $\mathcal{Y}$).

In our case, we define a context set $\mathcal{D}_c = \{\mathbf{o}_{c,n}, \mathbf{s}_{c,n}\}_n$ of observations $\mathbf{o}_{c,n}$ and *source labels* $\mathbf{s}_{c,n}$. Source labels can be, e.g., indicators for 'observational', 'interventional,' 'counterfactual,' etc., corresponding to each $I_k \in \mathbb{I}$. We then define the target set $\mathcal{D}_t = \{\mathbf{x}_{t,m}, \boldsymbol{\theta}_{t,m}\}_m$ consisting of m query points $\mathbf{x}_{t,m}$ and corresponding realized causal query values $\boldsymbol{\theta}_{t,m}$. Context and target sets $\{(\mathcal{D}_c, \mathcal{D}_t)\}$ are generated from a joint causal-query-mixture distribution directly following Definition 6. The model then does the following operations (see also Ashman et al. (2025, Fig. 4)).

$$\mathcal{D}_c = \{\mathbf{o}_{c,n}, \mathbf{s}_{c,n}\}_n \xrightarrow{\text{Encode, } e(\cdot)} \{\mathbf{z}_{c,n}\}_n \xrightarrow{\text{Process, } \rho(\cdot, \mathbf{x}_t)} \mathbf{z}_t \xrightarrow{\text{Decode, } d(\cdot)} p(\boldsymbol{\theta}_t \mid \mathcal{D}_c, \mathbf{x}_t)$$

In the experiments in this paper, we make use of two CNP variants based on the above setup. The first is a standard mean-pooled CNP, and the second is an attention-based variant utilizing Set Transformers Lee et al. (2019) (a TNP variant). C/TNPs often assume a Gaussian distribution over the target variable and thus output the target mean and variance (trained via, e.g., Gaussian Negative Log Likelihood Loss). In our causal inference setting, with treatment effect distributions that can be skewed, heavy-tailed, or multi-modal, we instead opt to implement quantile regression, similar to the Conditional Quantile Neural Process (CQNP) Mohseni et al. (2023). We take the desired quantile level $\tau \sim \mathcal{U}(0, 1)$ as an input during the forward pass. τ is embedded and concatenated with the aggregated context before decoding. We train via Huber loss over randomly sampled quantiles for each batch. This allows the decoder to act as a continuous quantile function $Q(\boldsymbol{\theta}_t \mid \mathcal{D}_c, \mathbf{x}_t, \tau)$. We refer to these two variants as Q-CNP and Q-TNP, if relevant, and otherwise just say 'the meta-model.'

PATE vs. CATE/ITE meta-estimation. Architecturally, the distinction between PATE and CATE/ITE meta-estimation reduces to the routing of the query points $\mathbf{x}_t$. For population-level estimands like PATE, the decoder exclusively uses the global context, whereas for individual-level estimands like CATE/ITE, the meta-model conditions on $\mathbf{x}_t$, processing query features with the global context either via direct concatenation (Q-CNP) or cross-attention (Q-TNP).

3.2 Posterior estimates of the probability of identifiability

A benefit of estimation with uncertainty in our context is the ability to, given a trained model, directly estimate the probability of identifiability at test-time. Algorithm 1 (Appendix C) describes this process in the case of quantile regression (applicable to Q-CNP and Q-TNP), but analogous procedures would work for any model supporting posterior uncertainty. For data we generate, a more direct way to

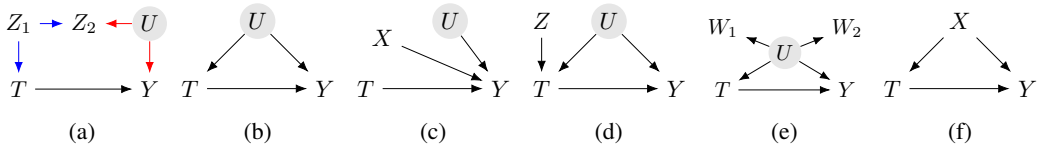


Figure 3: DAGs for each experimental setting we consider. **(a)** O-set Adjustment, **(b)** Hidden Confounder, **(c)** Hidden Moderator, **(d)** Instrumental Variable, **(e)** Proximal Case, **(f)** Known Confounder.

estimate computational identifiability is to directly compute the rate of predictions exceeding the desired error tolerance empirically over a test set. We do both in this paper. The benefit of theoretical estimates such as in Algorithm 1 is that we can in principle estimate computational identifiability for new data without ground truth effects. We leave a full characterization of this process to future work, but demonstrate how Algorithm 1 lines up with empirical estimates for in-meta-distribution samples in our experiments in Figure 5d. Figure 7 explains an example diagram of computational identifiability (the probability of identifiability) of the kind we compute throughout our experiments.

4 Experiments

In this section, we demonstrate computational identifiability across a variety of complex and small finite sample settings, including identification with ambiguous graphical criteria (Section 4.1), with mixed observational-interventional data (Section 4.2), and across counterfactual data and estimands (Section 4.3). Appendix D additionally includes experiments that illustrate how computational identifiability further generalizes other works on causal meta-prediction. Architectures, hyperparameter settings, and training details for each experiment can be found in Table 1 in Appendix E.

4.1 Optimal adjustment set selection

One use-case for computational identifiability is to distinguish between settings that are — from a theoretical identifiability perspective — all equally identifiable. A common setting for this is finding an *optimal valid adjustment set (O-set)* (Henckel et al., 2022) from among multiple valid sets, defined as a covariate adjustment set that provides optimal asymptotic variance for estimation.

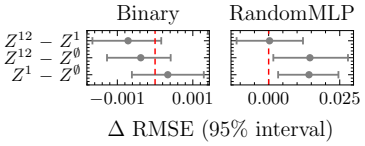
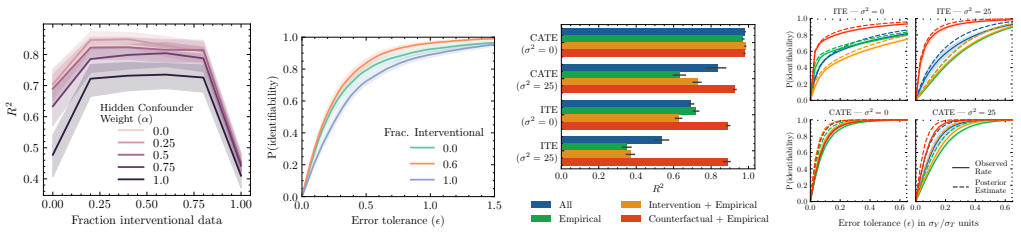


Figure 4: Differences in test RMSE under different adjustment sets

Works such as Henckel et al. (2022) provide graphical criteria for determining the O-set in linear causal models, extended to the non-parametric case in Rotnitzky & Smucler (2020). Rotnitzky & Smucler (2020) also discuss several cases where an optimal adjustment set does not exist according to graphical criteria. One such example is the directed acyclic graph (DAG) shown Figure 3a, with treatment T, outcome Y, and unobserved variable U. There are three valid adjustment sets: $\mathbf{Z}^0 = \emptyset$, $\mathbf{Z}^1 = \{Z_1\}$, and $\mathbf{Z}^{12} = \{Z_1, Z_2\}$. Rotnitzky & Smucler (2020) show that $\mathbf{Z}^0$ is uniformly better than $\mathbf{Z}^1$, but that determining which adjustment set is optimal between $\mathbf{Z}^0$ and $\mathbf{Z}^{12}$ would depend on the structural equations: $\mathbf{Z}^0$ is better than $\mathbf{Z}^{12}$ if the blue edges encode strong associations and the red edges are weak, while $\mathbf{Z}^{12}$ is better than $\mathbf{Z}^0$ if instead the red edges are strong and the blue edges are weak. Rotnitzky & Smucler (2020, Ex. 5) demonstrate this by showing the existence of two parameter settings of the following SCM: $U \sim \text{Bern}(p_u)$; $Z_1 \sim \text{Bern}(p_{z_1})$; $Y \sim \text{Bern}(p^Y_{11} UA + p^Y_{10} U\neg A + p^Y_{01} \neg UA + p^Y_{00} \neg U\neg A)$; $A \sim \text{Bern}(p^A_1 Z_1 + p^A_0 \neg Z_1)$; and $Z_2 \sim \text{Bern}(p^{Z_2}_{11} Z_1 U + p^{Z_2}_{10} Z_1 \neg U + p^{Z_2}_{01} \neg Z_1 U + p^{Z_2}_{00} \neg Z_1 \neg U)$ — one setting where $\mathbf{Z}^0$ is optimal and one where $\mathbf{Z}^{12}$ is optimal. But which adjustment set would be optimal in general, if, say, we had a prior over the parameters $\pi(p_j)$ and an observed dataset of size $n = 100$? Computational identifiability can answer this question directly: given a prior and hypothesis space of estimators, we simply meta-train an estimator and inspect which adjustment set leads to the lowest test error.

Figure 4 shows 95% confidence intervals of this comparison across 1000 test samples and five training runs. In the binary case, the prior corresponds to parameters $p_j \sim \mathcal{U}(0, 1)$ for all j. In this case, estimation is easy, there is no noise, and all models have comparable RMSE; however, we could still say $\mathbf{Z}^{12}$ is the *conditional O-set*. But what about in a case where estimation is difficult? The second



(a) (b) (c) (d)

Figure 5: (a) Meta-model performance in the transportability setting with varying mixtures of observational-interventional data. (b) Identifiability curves for the transportability setting. (c) Meta-model performance in the counterfactual setting estimating CATE vs. ITE. (d) Identifiability curves for the counterfactual setting, showing both empirical rates and posterior estimates from Algorithm 1.

column shows the case where instead the prior over SCMs is random multi-layer perceptrons (MLPs) (2 layers each) with Gaussian noise: $U \sim \mathcal{N}(0,1)$, $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 = \text{RandomMLP}(Z_1, U) + \mathcal{N}(0,1)$, $A = \text{RandomMLP}(Z_1) + \mathcal{N}(0,1)$, and $Y = \text{RandomMLP}(A, U) + \mathcal{N}(0,1)$. In this case, following the same empirical process, we can determine the conditional O-set is instead $\mathbf{Z}^{\emptyset}$.

4.2 Transportability with observational-interventional mixtures

Being able to transport causal effects across populations is a key concern for causal inference in practice. A common use-case for transportability is to be able to take information from a randomized control trial and use it to recover a treatment effect in an observational study. To explore identification in the transportability setting, we consider an example task of being able to identify a causal effect under hidden confounding (Figure 3b) with access to varying amounts of interventional data, where the source distribution of treatment T in the interventional data *does not match* the distribution in the observational data. Specifically, we consider the following prior over SCMs: $U_U, U_T, U_Y \sim \mathcal{N}(0,1)$; $U = U_U$; $T = f_t(\alpha U + U_T\sqrt{1-\alpha^2})$; $Y = f_y(T, \alpha U + U_Y\sqrt{1-\alpha^2})$; with random MLPs f_t, f_y and hidden confounder strength parameterized by α, such that we have constant variance $\mathbb{V}(\alpha U + U\sqrt{1-\alpha^2}) = \alpha^2\mathbb{V}(U) + (1-\alpha^2)\mathbb{V}(U) = 1$. We consider varying levels of hidden confounder strength as well as varying fractions of interventional vs. observational data available, where the interventional data come from a regime under $\text{do}(T \leftarrow \mathcal{U}(-1,1))$, but the target treatment effect to identify is the effect of a unit increase $\text{do}(T \leftarrow T+1)$ relative to the null-intervention.

Figure 5a shows R^2 of the meta-model in the transportability setting as hidden confounder weight α varies from 0 to 1 and the relative fraction of observational:interventional data varies from 0 to 1. Interventional data do indeed help identify the effect, with a sharp increase in performance, especially under stronger confounding. However, Figure 5a also shows a sharp decrease when all data are interventional, to the point where having all interventional data is worse than having none. This demonstrates the nuance of transporting the causal effect from an interventional regime whose treatment distribution does not directly reflect the target. Figure 5b summarizes this result using the identifiability curve for three fractions (0, 0.6, and 1), where uncertainty bands also represent the range of P(identifiability) across α values. In summary, computational identifiability has allowed us to glean an empirical answer for identifiability in this mixture setting and to learn something practical about the mix of observational-interventional data needed to maximize our chances of identifiability.

4.3 Counterfactual computational identification

Counterfactual reasoning is a core component of causal understanding, planning, and decision making. Theoretical counterfactual identification can be particularly challenging, given the need to model combined empirical and hypothetical data as well as introduce shape constraints or otherwise constrain the space of counterfactual distributions in order to reason and transfer information across counterfactual worlds (e.g., Correa et al., 2021; Maiti et al., 2025). In this experiment, we demonstrate computational identification in a small finite-sample counterfactual case. Specifically, we consider comparing identification of the CATE to that of the ITE for the DAG in Figure 3c with $\beta_X, \beta_T, \beta_{UT} \sim \mathcal{U}(-3,3)$, $\beta_{XT} \sim \mathcal{U}(-1,1)$; $X \sim \mathcal{U}(-2,2)$, $T \sim \text{Bern}(0.5)$, $\epsilon \sim \mathcal{N}(0,\sigma^2)$, and $Y = \beta_X X + \beta_T T +$

$\beta_{XT}XT + \beta_{UT}UT + \epsilon$. In this setting, there is no unobserved confounding, but we have unobserved variable U that moderates the treatment effect. This makes the CATE, which is an expected average treatment effect conditional on X, distinct from the ITE, which is instead a unit-specific treatment effect depending also on U. There is also noise ϵ on the outcome with variance σ^2. We consider a mix of empirical, interventional, and counterfactual data, where both interventions and counterfactuals correspond to do$(T \leftarrow (1 - T))$, creating a corresponding unit with the opposite treatment for each observed data point. Interventional pairs resample U and resample noise, but condition on X, while counterfactual pairs perform abduction (and thus resample neither U nor noise). Figure 5c shows test R^2 after meta-training in 16 different settings with $n = 100$ observations. We consider 100% empirical data (Empirical), 50% empirical data with interventional pairs (Interventional + Empirical), 50% empirical data with counterfactual pairs (Counterfactual + Empirical), and an equal mix of all three types (All). We meta-train on each mixture for both the CATE and the ITE in a noiseless setting ($\sigma^2 = 0$) as well as a noisy setting ($\sigma^2 = 25$). Figure 5d shows corresponding P(identifiability) curves for each setting – this time showing both observed empirical rates (solid lines) and posterior estimates from Algorithm 1 (dashed lines).

Several takeaways about computational identifiability for the CATE and ITE are evident here. First, computational identification of the CATE works in all settings, requiring neither interventional nor counterfactual data. Second, both counterfactual and interventional data can, however, help to recover the CATE when the outcome is noisy. This is in line with variance reduction from, e.g., matched pairs and/or common random numbers. Third, unlike the CATE, computational identification of the ITE requires counterfactual data, and — as we might expect from the definition of abduction — works equally well regardless of outcome noise. Finally, the interventional data (Interventional + Empirical) *do not help* identify the ITE in this case, and if there is no noise on the outcome, *they can instead make things worse*, even under the same interventional regime.

Non-monotonic and architecture-specific identifiability. The counterfactual case above — and in particular two negative results — allow us to demonstrate an important point about computational identifiability and its conditional nature. They can be summarized intuitively as 'conditional really does mean conditional.' Two parameters of our prior and hypothesis space demonstrate this point well in the ITE case — dataset size n and our choice of architecture. Figure 6 shows that when we increase dataset size, estimation of the CATE remains the same or improves, but ITE estimation can be substantially worse. This demonstrates how computational identifiability can have non-monotonic or surprising relationships with things like increasing dataset size that contradict what we might normally expect from theoretical identifiability. A similar result occurs with architecture choice, in this case whether or not the meta-model processes query points and context sets via cross-attention and set attention (as in the Q-TNP) or via concatenation and mean-pooling (as in the Q-CNP). Taking the (Counterfactual + Empirical, $\sigma^2 = 25, n = 100$) case as an example, the Q-CNP was only able to achieve $R^2 = 0.4701$, while the Q-TNP was able to achieve $R^2 = 0.8931$. In other words, in this experiment, only the latter architecture was able to learn to perform the matching of individual datapoints that recovering the ITE required. This property of being conditional on an empirical search — and its chosen hyperparameters — is a core component of computational identifiability. The notion itself is fundamentally about attaching a practical, hypothesis-constrained search process to the concept of identifiability.

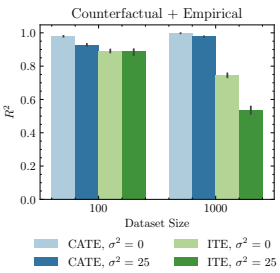


Figure 6: Non-monotonic relationship with size n

5 Conclusion and Future Work

Identifiability is not a notion specific only to causal inference, and there is potential to expand the set of definitions we have written in Section 3 to work beyond the causal inference settings to other settings, like survival analysis (Ranganath et al., 2016; Miscouridou et al., 2018) and missing data (Little & Rubin, 2019), where assumptions and analysis drive traditional theoretical identifiability. Another important future direction is to study computational identifiability in high dimensional settings, which we have not considered here. A key benefit of computational identifiability, as we have defined it, is that it makes the notion of identifiability *actionable*. Through additional optimization of the computational search process itself, there is also significant potential to further make such a search for identifiability systematic or automated.

AI USE STATEMENT

In this work, we used generative AI tools for some aspects of code editing, for example to assist with implementation of scripts for launching and managing SLURM jobs, as well as for suggestions to improve, format, and debug our codebase after its initial layout, structuring, and implementation. AI-assisted code examples were treated akin to personalized Stack Overflow recommendations: structured suggestions to use as a reference and help answer questions while developing, but not to use as-is without careful review or documentation. We have reviewed all AI-assisted work, for example, manually vetting suggested changes and reproducing detected bugs before implementing whichever fixes were deemed correct and appropriate. We take responsibility for the final content of this work, including any artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All experimental settings and implementation choices are described in detail in the paper and its appendix, and corresponding anonymized code for our framework and all experiments can be found at `https://anonymous.4open.science/r/compID-5B4E`. Detailed experiment descriptions can be found in Section 4, complete model architecture descriptions in Section 3.1.1, an illustration of how to map our framework and notation into a concrete implementation in Appendix D, and training procedures and hyperparameter choices for each experiment and figure in the paper in Table 1 in Appendix E.

REFERENCES

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

Matthew Ashman, Cristiana Diaconu, Eric Langezaal, Adrian Weller, and Richard E Turner. Gridded transformer neural processes for spatio-temporal data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 1722–1761. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/ashman25a.html`.

Yang Bai, Yifan Cui, and Baoluo Sun. Proximal inference for indirect and intervening effects in population interventions. *arXiv preprint arXiv:2504.11848*, 2025.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C Cresswell, and Rahul Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Vahid Balazadeh, Hamidreza Kamkari, Medha Barath, Ricardo Silva, and Rahul G Krishnan. Ivicl: Bounding causal effects with instrumental variables via in-context learning. *arXiv preprint arXiv:2605.12924*, 2026.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pp. 46–54. Elsevier, 1994.

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

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, 2012.

Bariscan Bozkurt, Houssam Zenati, Dimitri Meunier, Liyuan Xu, and Arthur Gretton. Density ratio-free doubly robust proxy causal learning. (arXiv:2505.19807), March 2026. doi: 10.48550/arXiv.2505.19807. URL `http://arxiv.org/abs/2505.19807`. arXiv:2505.19807 [cs.LG].

Lucius EJ Bynum, Aahlad Manas Puli, Diego Herrero-Quevedo, Nhi Nguyen, Carlos Fernandez-Granda, Kyunghyun Cho, and Rajesh Ranganath. Black box causal inference: Effect estimation via meta prediction. *arXiv preprint arXiv:2503.05985*, 2025.

Zhihong Cai, Manabu Kuroki, Judea Pearl, and Jin Tian. Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics*, 64(3):695–701, 2008.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions:causal effect identification and surrogate experiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06): 10093–10100, Apr. 2020a. doi: 10.1609/aaai.v34i06.6567. URL https://ojs.aaai.org/index.php/AAAI/article/view/6567.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020b.

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6856–6867. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/36bedb6eb7152f39b16328448942822b-Paper.pdf.

Anish Dhir, Cristiana Diaconu, Valentinian Mihai Lungu, James Requeima, Richard E Turner, and Mark van der Wilk. Estimating interventional distributions with uncertain causal graphs through meta-learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Oliver Dukes, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Proximal mediation analysis. *Biometrika*, 110(4):973–987, 2023.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981–995, 2007.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pp. 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

AmirEmad Ghassami, Ilya Shpitser, and Eric Tchetgen Tchetgen. Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*, 2023.

Helen Guo, Beatrix Yaxin Wen, and Ilya Shpitser. Proximal identification and estimation in front-door causal structures with unobserved confounding of the mediator. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 1857–1881. PMLR, 17–21 Aug 2026. URL https://proceedings.mlr.press/v337/guo26c.html.

James J Heckman and Edward J Vytlacil. Instrumental variables, selection models, and tight bounds on the average treatment effect. Working Paper 259, National Bureau of Economic Research, August 2000. URL http://www.nber.org/papers/t0259.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.

Joel L Horowitz and Charles F Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95(449): 77–84, 2000.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 217–224, 2006.

Emil Javurek, Dennis Frauen, Marie Brockschmidt, Jonas Schweisthal, and Stefan Feuerriegel. Amortizing causal sensitivity analysis via prior data-fitted networks. *arXiv preprint arXiv:2605.10590*, 2026.

Brendan Kline and Elie Tamer. Recent developments in partial identification. *Annual Review of Economics*, 15:125–150, 2023.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 06 2014. ISSN 0006-3444. doi: 10.1093/biomet/ast066. URL https://doi.org/10.1093/biomet/ast066.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Sanghack Lee and Elias Bareinboim. Causal effect identifiability under partial-observability. In *International Conference on Machine Learning*, pp. 5692–5701. PMLR, 2020.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. *arXiv preprint arXiv:2506.10914*, 2025.

Aurghya Maiti, Drago Plecko, and Elias Bareinboim. Counterfactual identification under monotonicity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26841–26850, 2025.

Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3080–3088. PMLR, 2019.

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

Xenia Miscouridou, Adler Perotte, Noémie Elhadad, and Rajesh Ranganath. Deep survival analysis: Nonparametrics and missingness. In *Machine Learning for Healthcare Conference*, pp. 244–256. PMLR, 2018.

Peiman Mohseni, Nick Duffield, Bani Mallick, and Arman Hasanzadeh. Adaptive conditional quantile neural processes. In *Uncertainty in Artificial Intelligence*, pp. 1445–1455. PMLR, 2023.

Francesco Montagna, Max Cairney-Leeming, Dhanya Sridhar, and Francesco Locatello. Demystifying amortized causal discovery with transformers. *arXiv preprint arXiv:2405.16924*, 2024.

Chan Park, David Richardson, and Eric Tchetgen Tchetgen. Single proxy control. *arXiv preprint arXiv:2302.06054*, 2023.

Judea Pearl. Causal diagrams for empirical research (with discussions). In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 255–316. 2022.

Judea Pearl and James M Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *UAI*, volume 95, pp. 444–453, 1995.

Aahlad Puli and Rajesh Ranganath. General control functions for causal effect estimation from ivs. *Advances in neural information processing systems*, 33:8440–8451, 2020.

Aahlad Puli, Adler Perotte, and Rajesh Ranganath. Causal estimation with functional confounders. *Advances in neural information processing systems*, 33:5115–5125, 2020.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Arvind Raghavan and Elias Bareinboim. Causal identification from counterfactual data: Completeness and bounding results. *arXiv preprint arXiv:2602.23541*, 2026.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pp. 101–114. PMLR, 2016.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188): 1–86, 2020. URL http://jmlr.org/papers/v21/19-1026.html.

Michael C Sachs, Gustav Jonzon, Arvid Sjölander, and Erin E Gabriel. A general method for deriving tight symbolic bounds on causal effects. *Journal of Computational and Graphical Statistics*, 32(2): 567–576, 2023.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

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

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *AAAI*, pp. 1219–1226, 2006.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Ilya Shpitser, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen. The proximal id algorithm. *Journal of Machine Learning Research*, 24(188):1–46, 2023.

James H Stock and Motohiro Yogo. Testing for weak instruments in linear iv regression, 2002.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Consistency of neural causal partial identification. *Advances in Neural Information Processing Systems*, 37:68956–68999, 2024.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Leslie G Valiant. A theory of the learnable. In *Foundations of Computation and Machine Learning: The Work of Leslie Valiant*, pp. 69–94. 2026.

Silvan Vollmer, Niklas Pfister, and Sebastian Weichwald. Identifying causal effects using a single proxy variable. (arXiv:2604.09135), April 2026. doi: 10.48550/arXiv.2604.09135. URL http://arxiv.org/abs/2604.09135. arXiv:2604.09135 [stat.ML].

Yong Wu, Yanwei Fu, Shouyan Wang, and Xinwei Sun. Doubly robust proximal causal learning for continuous treatments. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47233–47276, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ce530d24b0d26a5927534d0edc9d4ddd-Paper-Conference.pdf.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Kevin Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *Proceedings of the 11th Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=vouQcZS8KfW, 2022.

Liyuan Xu and Arthur Gretton. Kernel single proxy control for deterministic confounding. (arXiv:2308.04585), March 2025. doi: 10.48550/arXiv.2308.04585. URL http://arxiv.org/abs/2308.04585. arXiv:2308.04585 [stat.ML].

Lily Zhang, Veronica Tozzo, John Higgins, and Rajesh Ranganath. Set norm and equivariant skip connections: Putting the deep in deep sets. In *International Conference on Machine Learning*, pp. 26559–26574. PMLR, 2022.

A COMPLETE PRELIMINARIES: CAUSAL MODELS, INTERVENTIONS, AND COUNTERFACTUALS

Let capital letter X denote a random variable, where lowercase letter $X = x$ denotes the value it obtains. Let boldface capital letter $\mathbf{X} = \{X_1, \ldots, X_n\}$ denote a set of random variables, with value $\mathbf{X} = \mathbf{x}$. Capital P_X denotes the cumulative distribution function of variable X and lowercase p_X denotes the density (or mass) function. Let $P_{Y|X=x}$ denote the conditional distribution of Y given $X = x$ and $P_{Y|X}$ denote the collection of $P_{Y|X=x}$ for all x, i.e., the conditional of Y given X.

Causal models. We define a *causal model* as a tuple $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F})$. In this tuple, $\mathbf{V}$ is a set of observed variables, $\mathbf{U}$ a set of unobserved (exogenous) variables, and $\mathbf{F}$ a set of functions $\{f_i\}_{i=1}^{|\mathbf{V}|}$ for each $V_i \in \mathbf{V}$ such that $V_i = f_i(\text{PA}_i, U_i)$ where $\text{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$ represents the causal parents of V_i and $U_i \subseteq \mathbf{U}$. A causal model can be pictorially represented as a directed acyclic graph (DAG) with nodes for $\mathbf{U}, \mathbf{V}$ and directed edges for $\mathbf{F}$. A *probabilistic causal model* $(\mathcal{M}, P_{\mathbf{U}})$ adds distribution $P_{\mathbf{U}}$ over the unobserved variables. The 4-tuple $\mathfrak{C} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$ is also commonly referred to as a *SCM*. An SCM entails an *observational distribution* $P^{\mathfrak{C}}$ as well as distributions after interventions, which we define as follows.

Interventions. We define an *intervention* I on variable V_i as the substitution of equation $V_i = f_i(\text{PA}_i, U_i)$ with a modified mechanism $V_i := \tilde{f}_i(\tilde{\text{PA}}_i, \tilde{U}_i)$, commonly notated $\text{do}(I)$. Such 'soft' interventions Eberhardt & Scheines (2007) allow us to express several types of 'regimes' under which data are collected, for example allowing us to describe *transportability*, where we may wish to generalize causal knowledge across different environments with combined observational and/or experimental data Correa & Bareinboim (2020b). Soft interventions encompass the *idle* or *null* intervention $\emptyset$, which represents what happens naturally (i.e., $\tilde{f}_i = f_i$); *atomic* interventions, which set V_i to a constant value; *conditional* interventions, which set V_i to a deterministic function of observable parents; and *stochastic* interventions, which set V_i to a probability distribution conditional on a set of observable parents (see Correa & Bareinboim (2020b) for detailed discussion). A *submodel* $\mathcal{M}_I$ for intervention I is the model $\mathcal{M}$ after intervention $\text{do}(I)$. Given a probabilistic causal model, we can derive from $(\mathcal{M}_I, P_{\mathbf{U}})$ the distribution of any subset of variables following intervention $\text{do}(I)$. We denote an SCM $\mathfrak{C}$ after intervention I as $\mathfrak{C}^{\text{do}(I)}$, and the resulting *interventional distribution* as $P^{\mathfrak{C};\text{do}(I)}$. Note that $P^{\mathfrak{C};\text{do}(\emptyset)}$ is equivalent to the observational distribution $P^{\mathfrak{C}}$.

Counterfactuals. If, instead of using $P_{\mathbf{U}}$ in $(\mathcal{M}_{\mathbf{x}}, P_{\mathbf{U}})$ to derive an interventional distribution, we specify a distribution over the exogenous variables that is specific to a particular context or individual, the same mechanics that allow us to define interventions allow us to define *counterfactuals* that model alternate possible outcomes after interventions in a specific context. We use an asterisk to denote counterfactual versions $\mathbf{V}^*$ of variables $\mathbf{V}$. Counterfactual variable $\mathbf{Y}^*$ given a factual observation $\mathbf{z}$ and intervention I ($\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$) can be computed via a three-step procedure often referred to as 'abduction, action, prediction.' Abduction uses observed evidence to obtain $P_{\mathbf{U}|\mathbf{z}}$ from $P_{\mathbf{U}}$. Action performs intervention I to obtain $\mathcal{M}_I$. Prediction computes the probability of $\mathbf{Y}^*$ from $(\mathcal{M}_I, P_{\mathbf{U}|\mathbf{z}})$ (or analogously conditioning on $\mathbf{Z} \in A$ with $P(\mathbf{Z} \in A) > 0$ in the continuous case rather than on $\mathbf{Z} = \mathbf{z}$). We denote the *counterfactual distribution* $P^{\mathfrak{C}|\mathbf{Z}=\mathbf{z};\text{do}(I)}$.

B EXTENDED RELATED WORKS

Complete identification algorithms. Existing causal identifiability results and algorithms are sound and complete for certain types of models, such as recursive semi-Markovian SCMs. Do-calculus Pearl & Robins (1995); Pearl (2022) is a set of rules that systematically answer whether or not an interventional causal query is identifiable based on graphical criteria. Do-calculus has been show to be complete Huang & Valtorta (2006); Shpitser & Pearl (2006) for any query that can be expressed in terms of the do-operator. The ID and IDC algorithms Shpitser & Pearl (2008) applying do-calculus are complete identification algorithms for interventional as well as conditional interventional causal queries. To address more complex settings that cannot be expressed using the do-operator (such as path-specific effects and dynamic treatment regimes), po-calculus Malinsky et al. (2019) further generalizes do-calculus to provide identification for conditional path-specific

effects. For counterfactual queries, both unconditional and conditional, corresponding identification algorithms ID* and IDC* are complete Shpitser & Pearl (2008). Beyond hard (atomic) interventions on the target variable of interest, σ-calculus Correa & Bareinboim (2020a) inference rules provide complete identification with more general (soft) interventions that can be atomic, conditional, or stochastic (see Appendix A). Analogously, z-identifiability Bareinboim & Pearl (2012) describes identification of causal effects using surrogate experiments on variables other than the target treatment that are more accessible to manipulation in practice.

Bespoke and partial identification. In settings where general identification algorithms do not apply or where required identification assumptions are not met, novel conditions and algorithms are developed on a case-by-case basis to either point identify or bound target queries (e.g., Balke & Pearl, 1997; Horowitz & Manski, 2000; Cai et al., 2008; Shalit et al., 2017; Sachs et al., 2023). Common example settings with bespoke and/or parametric analysis include the instrumental variable setting Balke & Pearl (1994); Angrist et al. (1996); Heckman & Vytlacil (2000); Stock & Yogo (2002), the proximal setting Miao et al. (2018); Tchetgen et al. (2020); Ghassami et al. (2023), and counterfactual data settings (e.g., Raghavan & Bareinboim, 2026).

Neural Causal Models. A related line of work on neural causal models (NCMs) Xia et al. (2021) also connects identifiability and estimation, but in a fundamentally different manner than our work here. Neural identifiability Xia et al. (2022) shows that identification within the infinite space of possible NCMs is exactly equivalent to identification in SCM-space.[3] Unlike neural identifiability, the computational identifiability we propose here does not aim to search the SCM-space itself. Rather, computational identifiability aims to *directly learn a shortcut from data to estimator*, should such a shortcut exist. For this reason, even empirical approximations of neural identifiability are fundamentally distinct from computational identifiability.

Meta- and in-context learning for causal inference. A last connected line of research is that on meta-learning, in-context learning, and/or meta-prediction for causal effect estimation Bynum et al. (2025); Robertson et al.; Balazadeh et al.; Dhir et al.; Ma et al. (2025). Identification and sensitivity analysis in the meta-learning setting has only recently started to be considered Montagna et al. (2024); Balazadeh et al. (2026); Javurek et al. (2026). In this paper, we operationalize computational identifiability as an additional layer of abstraction on top of transformer-based meta-learning, taking causal meta-prediction methods as one sub-component of a larger identifiability pipeline. Importantly, any algorithm — neural-network-based or not, meta-learning-based or not — can in principle be incorporated into a computational identifiability search (see Appendix F for an example).

C Posterior estimates of identifiability

This section details the process by which we can, given a trained model, directly estimate the probability of identifiability at test-time. Algorithm 1 describes this process for meta-models that can serve as continuous quantile functions (such as the Q-CNP and Q-TNP), but analogous procedures would work for any model supporting posterior uncertainty. Figure 7 illustrates a *computational identifiability curve*, visualizing the probability of identifiability (i.e., the output of Algorithm 1) that estimates Equation (1) across a range of ϵ values.

[3] Neural identification algorithms (e.g., NeuralID Xia et al. (2022)) proceed by simultaneously maximizing and minimizing the causal query in NCM-space, and if the maximum and minimum query values are equivalent (via hypothesis test), identification is satisfied. The value of the query found during this process can then be used as an estimate.

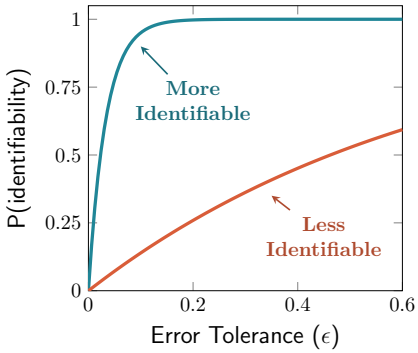


Figure 7: Diagram showing *computational identifiability curves*, visualizing Definition 4 across a range of possible ϵ values. The empirical (or posterior) probability of identifiability can be read at a given desired error tolerance.

Algorithm 1 Estimate P(identifiability) at ϵ via posterior uncertainty

Require: Test set $\mathcal{D}_{\text{test}}$, trained quantile function $\widehat{Q}$, dense quantile grid $\mathcal{T} = \{\tau_1, \ldots, \tau_K\}$, tolerance ϵ

1: **for** $(\mathbf{o}, \mathbf{x}) \in \mathcal{D}_{\text{test}}$ **do**

2: $\quad \phi \leftarrow \widehat{Q}_{0.5}(\mathbf{o}, \mathbf{x})$

3: $\quad \widehat{P}(\mathbf{o}, \mathbf{x}) \leftarrow \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left(\widehat{Q}_{\tau_k}(\mathbf{o}, \mathbf{x}) \in [\phi - \epsilon, \phi + \epsilon]\right)$ $\quad \triangleright \approx \int_0^1 \mathbf{1}\left(\left|\widehat{Q}_\tau(\mathbf{o}, \mathbf{x}) - \phi(\mathbf{o}, \mathbf{x})\right| \leq \epsilon\right) d\tau$

4: **end for**

5: $\widehat{P}_{\text{avg}} \leftarrow \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\mathbf{o},\mathbf{x}) \in \mathcal{D}_{\text{test}}} \widehat{P}(\mathbf{o}, \mathbf{x})$ $\quad \triangleright \approx \mathbb{E}_{\mathfrak{C} \sim \pi}\left[\mathbb{E}_{\mathbf{O} \sim p^{\mathfrak{C}}_{\mathbf{O},mix}}\left[\mathbb{E}_{\mathbf{X} \sim p_{query}}\left[\widehat{P}(\mathbf{O}, \mathbf{X})\right]\right]\right]$

6: **return** $\widehat{P}_{\text{avg}}$

D FURTHER GENERALIZING CAUSAL META-PREDICTION FRAMEWORKS

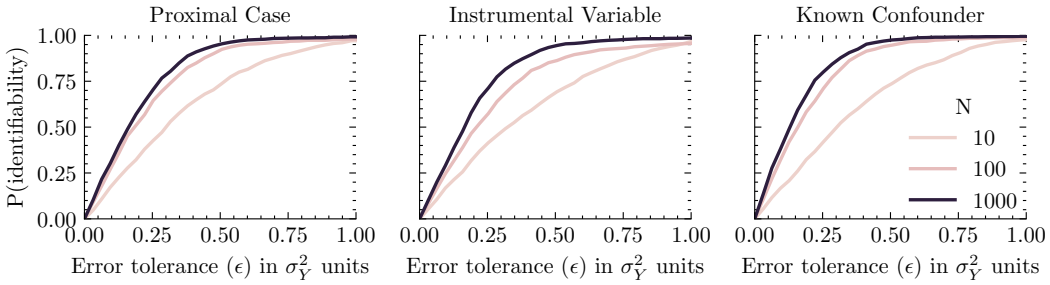


Figure 8: Identifiability curves from meta-models trained on each of the cases in Equation (2), as a function of varying dataset size n.

Works such as Bynum et al. (2025) have proposed frameworks to automate the process of finding estimators via meta-learning. Our abstractions in Section 3 further generalize this process, taking the meta-prediction of a causal estimator as one piece of a larger pipeline. To provide some intuition for how meta-prediction of effect estimators fits into our pipeline, we illustrate in this section how experiments from Bynum et al. (2025) are equivalent to a special case of our framework and notation.

Bynum et al. (2025, Sec. 3) introduce the following data generating process in Equation (2) to demonstrate meta-learning causal effect estimators in three different settings: a proximal causal inference case (Figure 3e), an instrumental variable case (Figure 3d), and a known confounder case

(Figure 3f).

$$\begin{aligned} \delta_1, \delta_2, \delta_3 &\sim \mathcal{U}(0,1) \\ \gamma_t, \beta_t, \beta_y &\sim \mathcal{U}(-1,1) \\ \gamma_x, \beta_x &\sim \mathcal{U}(-2,-1) \cup \mathcal{U}(1,2) \end{aligned} \qquad \begin{aligned} U_Y, Z, X &\sim \mathcal{N}(0,1) \\ W_1 &\sim \mathcal{N}(X, \delta_1) \qquad T = \gamma_x X + \gamma_t Z \\ W_2 &\sim \mathcal{N}(X, \delta_2) \qquad Y = \beta_x X + \beta_t T + \beta_y U_Y \end{aligned} \tag{2}$$

In the proxy case, W_1, W_2, T, Y are observed; in the instrument case, Z, T, Y are observed; and in the known confounder case, X, T, Y are observed.

In our framework, we can represent these cases via a simple causal mixture distribution, where most of the terms are set to zero ($\alpha_{ij}^{emp} = 0, \alpha_{ijk}^{cf} = 0, \mathbb{I}_{\mathbb{D}} = \emptyset$), and we have $\alpha_0^{int} = 1$ for the sole element of $\mathbb{I} = \{\emptyset\}$ — the null intervention. This leaves us with the following for $p^{\mathfrak{C}}_{\mathbf{O},mix}$ from Definition 2:

$$\begin{aligned} p^{\mathfrak{C}}_{\mathbf{O},mix}(\mathbf{o}) &= \sum_{I_j \in \mathbb{I}_{\mathbb{D}}} \sum_{\mathbf{o}_i \in \mathcal{D}_j} \left(\alpha_{ij}^{emp} \delta(\mathbf{o} - \mathbf{o}_i) + \sum_{I_k \in \mathbb{I}} \alpha_{ijk}^{cf} p^{\mathfrak{C}^{\text{do}(I_j)} | \mathbf{O} = \mathbf{o}_i; \text{do}(I_k)}_{\mathbf{O}}(\mathbf{o}) \right) \\ &\quad + \sum_{I_k \in \mathbb{I}} \alpha_k^{int} p^{\mathfrak{C}; \text{do}(I_k)}_{\mathbf{O}}(\mathbf{o}) \\ &= \sum_{I_k \in \mathbb{I}} \alpha_k^{int} p^{\mathfrak{C}; \text{do}(I_k)}_{\mathbf{O}}(\mathbf{o}) \\ &= p^{\mathfrak{C}; \text{do}(\emptyset)}_{\mathbf{O}}(\mathbf{o}) \\ &= p^{\mathfrak{C}}_{\mathbf{O}}(\mathbf{o}) \end{aligned}$$

or simply in this case, the observational distribution. In each case, we would have $\mathbf{O} = \{W_1, W_2, T, Y\}$, $\mathbf{O} = \{Z, T, Y\}$, and $\mathbf{O} = \{X, T, Y\}$, respectively.

We then define the following joint causal-query-mixture distribution according to Definition 6, for prior over SCMs π defined by Equation (2), and using the PATE as our causal query for intervention $\text{do}(T \leftarrow T + 1)$ relative to the null-intervention $\text{do}(\emptyset)$:

$$\begin{aligned} p_{joint}(\mathbf{o}, \mathbf{x}, \theta) &= \mathbb{E}_{\mathfrak{C} \sim \pi} \left[p^{\mathfrak{C}}_{\mathbf{O},mix}(\mathbf{o}) \cdot p_{query}(\mathbf{x} \mid \mathfrak{C}, \mathbf{o}) \cdot \delta \left(\theta - \Theta \left[\mathfrak{C}, \mathbf{x} \right] \right) \right] \\ &= \mathbb{E}_{\mathfrak{C} \sim \pi} \left[p^{\mathfrak{C}}_{\mathbf{O}}(\mathbf{o}) \cdot \delta \left(\theta - \Theta_{\text{PATE}} \left[\mathfrak{C}, \cdot \right] \right) \right] \\ &= \mathbb{E}_{\mathfrak{C} \sim \pi} \left[p^{\mathfrak{C}}_{\mathbf{O}}(\mathbf{o}) \cdot \delta \left(\theta - \mathbb{E}_T \left[\mathbb{E}_{p_Y^{\mathfrak{C}; \text{do}(T \leftarrow T+1)}}[Y] - \mathbb{E}_{p_Y^{\mathfrak{C}; \text{do}(\emptyset)}}[Y] \right] \right) \right] \end{aligned}$$

where query points $\mathbf{x}$ are not needed, since the PATE is a population-level query (see Definition 1 and the discussion thereafter).

Having defined a joint causal-query-mixture from which to stream training data $\{(\mathbf{o}, \theta)\}$, we then meta-train a Q-CNP (see Section 3.1.1) to perform effect estimation in each of the three cases and compute the empirical probability of identifiability over a new test set sampled from the same distribution.

Figure 8 shows the results of meta-training in each of these settings and the resultant identifiability curves as a function of one possible parameter of the computational search: dataset size n. We see (1) that meta-training is indeed successful in each of the cases and (2) that as dataset size increases, so too does the probability of identifiability. In this example, we see an illustration of how computational identifiability can allow us to concretely explore identification even with finite sample sizes as small as $N = 100$ or $N = 10$.

E Experiment details

This section details the experiment architectures, training regimes, and hyperparameters for all of the experiments in this paper. In order to more easily understand how our unified notation in Section 3 maps onto the process of meta-training transformers to learn causal estimators, we provide a step-by-step illustration of this mapping to a special case in Appendix D. Additionally, full descriptions of the Q-CNP and Q-TNP architectures can be found in Section 3.1.1. While several experiments

train multiple models, each individual meta-training instance used at most 1 GPU with at most 32GB memory for a compute time of up to 5 hours. Experiments were run on a mix of Intel Xeon Platinum 8592+ 64C CPUs, NVIDIA H200 GPUs, NVIDIA L40S GPUs, and an M5 MacBook Pro.

Bootstrap intervals. We detail here our procedure for computing bootstrap intervals across our experiments, which by default uses paired bootstrap samples across coupled sets of test tasks. Among our experiments, the proximal front-door experiments (Section 2 and Appendix E.1) and the optimal adjustment set experiments (Section 4.1) both draw comparisons across fully coupled test sets, as these are both cases where we consider different access to observed variables on what are otherwise identical tasks. In our other experiments (Sections 4.2 and 4.3 and Appendix D), test tasks in different cases consist of either different estimands, different datasets, or both, but may share some other forms of randomness, such as common random seeds for some parts of data generation before the settings diverge. Let $\hat{\theta}_{s,k}$ be the prediction of meta-model s on test task k, with true parameter value θ_k (which can be vector-valued). Consider a direct comparison of statistic T in Setting 1 vs. Setting 2 across N_{test} test tasks. Assume in each setting we have S trained models. For each of B iterations, we draw one bootstrap sample of the N_{test} task indices (paired across settings), and we independently draw two bootstrap samples of the S model indices (separate for each setting). We then recompute T for each setting as a function of the pooled $\{(\hat{\theta}_{s,k}, \theta_k)\}$ pairs over the $s = 1, \ldots, S$ resampled models and $k = 1, \ldots, N_{\text{test}}$ resampled tasks. This procedure produces B samples of T (or of a difference $T_1 - T_2$), from which we can report a 95% percentile bootstrap interval that reflects variation both across test tasks as well as across possible estimators. Increasing B is relatively inexpensive, but increasing S requires meta-training more models and is thus also a question of how much computation we want to spend on each experiment. When $S = 1$, the corresponding intervals report task variability only. We report values for each in Table 1.

Table 1: Table of meta-model architectures, training details (dataset size, number of query points, training regime (e.g., offline with a fixed number of examples or online streaming with fresh examples per epoch), and size of the test set), as well as hyperparameters (max epochs e, embedding size d, layers m, attention heads a, inducing points i, and number of quantiles sampled per batch q) for each experiment in the paper.

Experiment	Meta-model Architecture	Dataset Size	Query Points	Train Mode (Train Size)	Test Size	B, S	Hyperparameters					
							e	d	m	a	i	q
Figure 2c	Q-TNP	1000	-	Offline (40k)	1000	1000, 3	600	128	2	4	32	32
Figure 4	Q-CNP	100	-	Offline (100k)	1000	1000, 5	500	128	2	-	-	32
Figures 5a and 5b	Q-CNP	1000	-	Offline (10k)	1000	1000, 3	500	128	8	-	-	32
Figures 5c and 5d	Q-TNP	100	30	Offline (10k)	1000	1000, 5	500	128	2	4	32	32
ITE Architecture	Q-CNP/Q-TNP	100	30	Offline (10k)	1000	1000, 5	500	128	2	-/4	-/32	32
Figure 6	Q-TNP	100/1000	30	Offline (10k)	1000	1000, 5	500	128	2	4	32	32
Figure 8	Q-CNP	10/100/1000	-	Offline (40k)	1000	1000, 1	500	128	8	-	-	32
Figure 9	-	10/100/1000	-	-	1000	-	-	-	-	-	-	

E.1 Proximal metadentifiability of confounded front-door causal structures

This section provides the details of the data generating process for the proximal front-door experiment discussed in Section 2 and Figure 2c. Section 2 specifies the following meta-prior over the DAG in Figure 2a, with random linear $\{T, M\}$, random nonlinear $\{Z, W, Y\}$, and Gaussian $\{U_1, U_2\}$:

$$U_1, U_2 \sim \mathcal{N}(0,1) \qquad Z = f_z(U_1) + \epsilon_z \qquad M = f_m(U_1, T) + \epsilon_m$$
$$W = f_w(U_1) + \epsilon_w \qquad T = f_t(U_1, U_2) + \epsilon_t \qquad Y = f_y(U_1, U_2, M) + \epsilon_y$$

where f_t, f_m are random linear functions with uniform weights in $\mathcal{U}(-1, 1)$; f_z, f_w, f_y are random multi-layer perceptrons with 2 layers each; and $\epsilon_z, \epsilon_w, \epsilon_t, \epsilon_m, \epsilon_y \sim \mathcal{N}(0, 1)$.

F COMPUTATIONAL IDENTIFIABILITY WITH PER-DATASET ESTIMATION METHODS

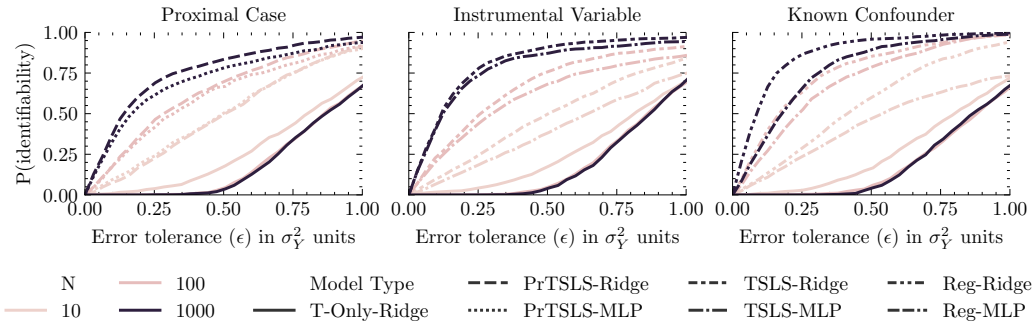


Figure 9: Identifiability curves from per-dataset estimation methods instead of meta-trained models, for each of the cases in Equation (2), as a function of varying dataset size n.

Using the example in Appendix D, we can demonstrate a useful point about computational identifiability; namely, that *the notion itself does not have to require meta-training*, and we can incorporate existing per-dataset algorithms into a computational identifiability search. Figure 9 shows exactly this, showing the probability of identifiability computed by empirically evaluating each per-dataset method described in Bynum et al. (2025, Sec. 3): proximal two-stage least squares with ridge regression (PrTSLS-Ridge), proximal two-stage least squares with an MLP first stage and ridge regression second stage (PrTSLS-MLP), two-stage least squares with ridge regression (TSLS-Ridge), two-stage least squares with an MLP first stage and ridge regression second stage (TSLS-MLP), regular ridge regression (Reg-Ridge), regression with an MLP (Reg-MLP), and finally ridge regression using treatment only (T-Only-Ridge).

This evaluation is done on each of the same datasets the meta-model in Figure 8 was evaluated on. However, comparing the superior performance of the meta-model to that of the baselines (especially those that perform poorly) illustrates the utility of having a meta-model even in settings with known algorithms, provided the meta-model is able to perform well. The story about identifiability in the meta-model search is, for example, more optimistic than in the known-algorithm search, especially for small dataset sizes.