Towards Trustworthy AI: From Local Explanations to Causal Understanding

Lucile Ter-Minassian

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2025

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the content of this thesis is my own work and has not been submitted in whole or in parts for any other degree or qualification. This thesis is my own work unless otherwise stated in the authorship form at the end of the chapters.

Lucile Ter-Minassian
Hilary 2025

I dedicate this thesis to all the women in science, and particularly to my grandmothers, Liliane and Lisbeth. Your thirst for knowledge has inspired my own, and this work stands on the shoulders of your courage and determination.

Acknowledgements

I am deeply grateful to my supervisor, Professor Chris Holmes, for his unwavering support and the intellectual freedom he has granted me throughout my DPhil. Despite the distance, your guidance, rooted in trust, has allowed me to explore my ideas with true independence while always feeling supported by a steady and thoughtful mentor. My gratitude also extends to Dr. Aisha Walcott Bryant, whose inspiration has stayed with me since our first meeting in Nairobi, long before this thesis began. Your passion for using science as a force for social good has profoundly shaped my own commitment to impactful research.

It took a village to get to where I am today, and I am profoundly thankful to all the professors and mentors who have believed in me and encouraged me along the way. I want to express my sincere appreciation to my internship mentors, particularly Liran, Ehud, and Yishai. Your generosity in welcoming me into your workplace, coupled with your mentorship, made my experience both intellectually enriching and truly enjoyable. These memories will remain treasured for years to come. To my collaborators – Sahra, Oscar, Jake, Desi, Robin, Uri and Karla – thank you for the insight, dedication, and teamwork. Working alongside you, especially through challenging deadlines, has been an invaluable learning experience.

A doctoral thesis can only flourish in the right environment. Marc-tonton, thank you for opening your home to me. Despite the rumble of the storms, Yafo's gentle birdsong and breathtaking sunsets have been my daily reminder that beauty persists. To my friends across the Mediterranean – you transformed strangers into family when the world seemed to be falling apart. Your resilience, laughter, and unfailing support have forever changed me. Thank you for welcoming me into your world with such warmth and grace.

To those who have walked alongside me for many years now: my forever friends and my siblings, thank you for the laughter, the deep conversations, and the unwavering support throughout these four years. Gabriel, Adele, Lazare, Emeraude, Clémence and so many others - I've drawn immense strength from your proud gazes and encouraging words.

To Shahine – la cabane – you've been my rock throughout my Oxford journey. Last year, you pulled me out of the darkness and guided my way back to my research when I needed it the most. Thank you for being such an amazing friend.

To my parents – your gentleness and your unwavering belief in me have sustained me. Papa, thank you for transmitting your determination and drive, and for being my motivation to pursue a scientific career. Maman, thank you for all our long conversations, but most of all, thank you for being the caring, funny, sensitive mother you are. I'm incredibly fortunate to have you both as parents. I love you deeply.

Finally, as I reflect on my time as a DPhil student, I cannot help but acknowledge how the world has changed in inexplicable ways. Paradoxically, while I immersed myself in the study of statistical causality and explainability, the reality around us seemed to follow patterns that escaped all logic and reason. The stark contrast between my academic pursuit of explanation and the inexplicable nature of global events has only deepened my commitment to science as a tool for clarity in an increasingly complex universe.

Most importantly, these experiences have made me profoundly aware of my privilege to question and learn - to pursue intellectual curiosity freely, to access education without barriers, and to dedicate years to deepening my understanding in an environment that nurtures critical thinking.

This thesis thus stands as a testament to the value of education, and in memory of those who can no longer walk the paths of learning.

Abstract

In an era of rapidly advancing AI capabilities and mounting ethical scrutiny, the demand for transparent and trustworthy machine learning solutions has never been more urgent. High-stakes domains such as healthcare, criminal justice, and financial services increasingly deploy complex AI systems that can profoundly impact human lives, yet their opaque nature often renders their decisions inscrutable to the very individuals affected by them. This lack of transparency not only raises ethical concerns but also poses practical challenges for practitioners who must ensure these systems behave as intended when deployed in real-world settings.

This dissertation explores the challenge of approximating different types of truth in machine learning, offering methodological improvements that advance both local explainability and causal inference. While XAI methods seek to illuminate the inner workings of black-box models, causal inference techniques aim to uncover the underlying mechanisms that drive observed phenomena. Together, these approaches form a comprehensive framework for understanding - XAI explains our models of reality, while causal inference explains reality itself. Throughout this work, I focus on improving approximations of various truth concepts: causal truths, model truths, and true data distributions.

My research begins by addressing foundational limitations in local model explanations, introducing Neighbourhood SHAP values that leverage local reference distributions to provide more meaningful feature attributions that better reflect local model behavior while demonstrating increased robustness against adversarial classifiers. Building on this foundation, I develop Path-Wise Shapley effects (PWSHAP), bridging predictive modeling and causal understanding by combining user-defined causal structures with Shapley values to assess variable impacts through specific causal pathways. Moving beyond model explanations to direct causal inference, I introduce BICauseTree, an interpretable balancing method that identifies clusters where natural experiments occur locally, detecting subgroups with positivity violations while providing explicit definitions of valid inference populations. Finally, I address the challenge of securely evaluating dataset value for causal investigations, developing an Expected Information Gain framework that allows data hosts to assess merge potential without compromising sensitive information. Collectively, these methodological innovations advance both theoretical understanding and practical applications in explainable and causal AI, offering solutions that enhance transparency, improve decision support, and strengthen trust in AI systems deployed in high-stakes domains.

Contents

Prologue

My journey into statistics was driven by a profound admiration for how statisticians help us understand the world through data. At its core, statistics is about inferring patterns from limited, observed data by fitting models that can hopefully approximate the underlying truth. This pursuit of data-driven insights naturally led me to explore two complementary approaches to answering "why": causal inference and explainable Artificial Intelligence (XAI).

As I delved deeper into these fields, I began to see them as different facets of the same fundamental challenge - approximating truth through data. Throughout this thesis, our focus will be on improving these approximations across various truth concepts: causal truths, model truths, and true data distributions. Causal inference transcends traditional correlation, allowing us to decipher the mechanisms behind observed phenomena. As Pearl and Mackenzie eloquently demonstrate in "The Book of Why," this approach empowers us to address the fundamental "why did X cause Y?" questions that drive scientific discovery [2].

Similarly, XAI develops methods to interpret why increasingly complex models make specific predictions. While causal inference explains phenomena in the world, XAI explains the reasoning within our models. Together, these fields form a comprehensive framework for understanding — causal inference explains reality, while XAI explains our models of reality.

As AI systems become more deeply integrated into critical domains, from healthcare to policy, the need for interpretability and accountability becomes paramount. Both causal inference and XAI contribute to making AI systems more transparent, robust, and ultimately more trustworthy. Causal inference ensures that decisions are grounded in real-world mechanisms, rather than mere statistical associations. XAI enables us to detect failure modes, understand model behavior, and communicate predictions in human-understandable terms. This dual pursuit—of external causal validity and internal model transparency—is at the heart of trustworthy AI. It is also the guiding thread of this thesis, as reflected in its title: *Towards Trustworthy AI: From Local Explanations to Causal Understanding.*

This thesis follows the format of an integrated thesis and is composed of 7 chapters. The first, second, and last chapters consist of an introduction, background, and conclusion respectively. The remaining 4 chapters are formed of separately published papers, which each contain their own introduction and a literature review specific to the topics covered therein. In this introduction, we would like to give a more general overview and motivation of the work done throughout the DPhil.

Chapter 1

Introduction

1.1 Contextual Motivations

1.1.1 Causal Inference

The limitations of predictive modelling approaches are increasingly evident across high-stakes domains. Predictive models often fail to address fundamental interventional questions: What are the consequences of altering a specific variable? How do various factors causally interact? Causal inference enables robust decision-making by identifying the underlying mechanisms that drive observed outcomes, thus facilitating targeted interventions.

Rubin's potential outcomes framework and Pearl's structural causal models have revolutionised our understanding of causal inference, providing rigorous methodological foundations for moving beyond associational reasoning [3, 4]. By leveraging causal inference, researchers can more accurately assess the true impact of interventions, distinguishing between mere correlation and genuine causal mechanisms [5, 6]. Demonstrating causal links is thus crucial in healthcare and policy-making, to ensure evidence-based decision-making.

Historical Context and Evolution

Causal inference has roots in multiple disciplines, evolving from early experimental design principles in agriculture to sophisticated frameworks in statistics, econometrics, and computer science. The formalization of causal reasoning through Rubin's potential outcomes framework and Pearl's structural causal models marked pivotal developments that transformed how researchers approach cause-effect relationships [3, 4]. While foundational frameworks have advanced causal reasoning, their application to complex, real-world data presents methodological challenges that underscore the need for refined causal tools.

The Need for Causal Reasoning in Complex Systems

In applied settings, individuals or groups receiving an intervention—whether medical treatment, policy exposure, or educational programs—often differ systematically from those who do not. Such baseline differences introduce selection bias, confounding treatment effects with pre-existing characteristics. Observational data reflects this entanglement, complicating efforts to isolate causal effects. Causal inference methodologies address these challenges, enabling researchers to disentangle the true impact of interventions from background variation.

Healthcare Imperatives
In medical research, understanding causal pathways directly impacts patient outcomes. When evaluating treatment options, the critical question concerns causal effects rather than mere statistical associations. Without this distinction, clinical decisions risk being influenced by spurious correlations. For instance, observational data might suggest improved outcomes among patients using a particular medication, yet this association may reflect underlying patient characteristics associated with negative outcomes, rather than treatment efficacy [5].

Policy Evaluation Challenges
Policymakers require causal knowledge to allocate resources effectively and design impactful interventions. Economic and social programs often involve recursive socio-economic dynamics that simple correlational approaches cannot disentangle. The evaluation of a job training program, for example, necessitates accounting for confounders such as participant motivation and prior skills, which can bias naive associations between program participation and employment outcomes [7].

From Association to Causation: Real-World Applications

The versatility of causal inference methodologies is evident across diverse domains, where establishing causality is critical for informed policy and practice:

- Public health interventions: Evaluating vaccination programs, smoking cessation initiatives, and nutritional guidelines requires understanding causal pathways to distinguish effective interventions from confounding factors [5].
- Economic policy: Central banks and regulatory agencies increasingly rely on causal analysis to predict the effects of interest rate changes, tax reforms, and market regulations [7].

- Environmental management: Robust causal analysis is essential for attributing environmental changes to human activities versus natural variation, guiding climate policy and conservation efforts [8].
- Education policy: Recent studies utilizing causal inference have assessed the impact of remote learning interventions on academic outcomes during the COVID-19 pandemic [9].
- Healthcare resource allocation: Causal models have been instrumental in optimizing hospital resource distribution during public health crises [10].

1.1.2 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence encompasses methods and techniques that help humans understand how AI systems arrive at their decisions. While traditional AI approaches may prioritize predictive performance, XAI provides interpretations of model predictions, highlights influential features, and offers insights into decision boundaries. These explanations serve various stakeholders - from model developers and domain experts to end-users and regulators - each requiring different levels of technical detail and contextual relevance when interacting with AI systems.

Historical Context and Regulatory Evolution

The quest for transparency in algorithmic decision-making is not a recent phenomenon. In the 1970s, St. George's Hospital Medical School (United Kingdom) implemented an admissions algorithm that was later revealed to discriminate against women and individuals with non-European names [11]. This historical precedent highlights that algorithmic opacity and its consequences have been concerns for decades, though they have gained unprecedented attention in recent years.

Modern explainable AI (XAI) research has accelerated dramatically as AI systems increasingly influence critical aspects of society. This growth has been further catalyzed by institutional and regulatory responses. The European Union's General Data Protection Regulation (GDPR) marked a pivotal turning point by introducing a "right to explanation" for algorithmic decisions that "significantly affect" individuals [12]. This legislative framework requires that algorithmic results be traceable and comprehensible upon request, transforming transparency from a technical preference to a legal obligation.

In parallel, specialised regulatory frameworks have emerged in high-stakes domains. The U.S. Food and Drug Administration (FDA) has developed guidance specifically addressing

AI-based medical devices, emphasizing transparency and interpretability as essential components of safety evaluation [13]. These regulatory developments represent a significant shift in how AI systems are developed, deployed, and governed.

Core Motivations for Explainable AI

Ensuring Fairness and Accountability Machine learning systems trained on historical data risk perpetuating and amplifying existing societal biases. *Obermeyer et al.* demonstrated this concern in their analysis of algorithms that systematically disadvantaged certain demographic groups due to proxy variables that inadvertently encoded bias [14]. Similar issues have been documented across domains ranging from judicial decision-making [15] to facial recognition [16]. Without transparent mechanisms to interrogate model decisions, such discriminatory patterns can remain concealed within the complexity of modern AI systems. XAI techniques enable developers and stakeholders to assess whether models make equitable decisions across demographic groups by providing visibility into the features and patterns driving predictions [17]. This transparency is crucial for identifying and mitigating potential sources of bias before deployment in sensitive contexts. *Facilitating Trust in Critical Applications* As AI systems increasingly inform high-stakes decisions, the need for interpretability becomes paramount [18]. Opaque algorithmic decisions can erode trust among both domain experts and end-users, particularly when errors occur without clear explanation. Research by Kizilcec [19] demonstrates that appropriate explanations significantly increase user trust and acceptance of algorithmic decisions. *Supporting Regulatory Compliance* Emerging regulatory frameworks increasingly mandate some form of algorithmic transparency [20]. The European Union's General Data Protection Regulation (GDPR), for instance, establishes a "right to explanation" for automated decisions with significant effects [21]. XAI methods provide practical means to satisfy these requirements while maintaining algorithmic performance.

Operational Benefits of XAI

- **Error Detection and Debugging:** Transparent models allow domain experts to identify when an AI system operates outside its validated domain or bases recommendations on spurious correlations [22].
- **Knowledge Validation:** Explanations that align with domain expertise enhance confidence in AI systems and facilitate appropriate reliance on algorithmic recommendations [23].
- **Stakeholder Communication:** In discussing AI-informed decisions, comprehensible explanations enable effective communication with diverse stakeholders who may lack technical expertise [24].

- **Feature Engineering:** XAI methods that quantify feature importance facilitate the development of more efficient, parsimonious models by identifying the most predictive variables [25].

Practical Applications Across Domains

Mitigating Representation Disparities Explainability techniques can reveal whether models perform consistently across different subpopulations [26]. This visibility is crucial, as data often suffers from representation disparities, missing information patterns correlated with demographic factors, and historical biases in collection procedures [11]. Beyond data issues, model architecture choices, optimization functions, and regularization techniques can introduce additional implicit biases. *Developing Interpretable Decision Aids* In resource-constrained environments, simpler models often prove more practical [27]. XAI methods that quantify feature importance facilitate the development of streamlined decision aids. These parsimonious models transform complex algorithms into practical tools that balance accuracy with usability [28, 29]. *Enabling Actionable Insights* When AI systems influence important recommendations, all stakeholders deserve to understand the rationale. Explanations that identify key factors can convert opaque predictions into actionable insights, empowering users to make informed decisions [30, 31]. This aspect of explainability transforms AI from a black-box oracle into a tool for education and engagement.

1.1.3 Balancing Complexity and Interpretability

The advancement of XAI techniques represents a crucial counterbalance to the increasing complexity of modern machine learning systems. As models grow more sophisticated, the gap between algorithmic capacity and human understanding widens [32]. Explainable AI serves as a bridge across this divide, ensuring that as we harness more powerful computational techniques, we maintain the ability to scrutinize, validate, and justify the resulting decisions. This balance is essential for fostering trust and facilitating the integration of AI into high-stakes domains [33].

1.2 Thesis Outline

This thesis contains four published papers followed up with a conclusion.

On Locality of Local Explanation Models

Sahra Ghalebikesabi*, **Lucile Ter-Minassian***, Karla Diaz-Ordaz, Chris Holmes.
Advances in Neural Information Processing Systems Conference (NeurIPS), 2021.

In Chapter 3, we investigate the extent to which Shapley values truly capture local model behaviour. We highlight limitations in conventional approaches that rely on global population distributions and introduce Neighbourhood SHAP values, which leverage local reference distributions to enhance local interpretability. This formulation improves on-manifold explainability, provides more meaningful feature attributions, and increases robustness against adversarial classifiers.

PWSHAP: a path-wise explanation model for targeted variables

Lucile Ter-Minassian*, Oscar Clivio*, Karla Diaz-Ordaz, Robin Evans, Chris Holmes.
International Conference on Machine Learning (ICML), 2023.

In Chapter 4, we explore how to interpret complex predictive models in sensitive domains where certain variables, like treatment effects in clinical settings or protected attributes in policy models, are of particular interest. We introduce Path-Wise Shapley effects (PWSHAP), a framework that combines user-defined causal structures with Shapley values to assess the targeted impact of binary variables. This approach enables clearer local bias and mediation analyses, supports the quantification of local effect modification when variables are randomised, and strengthens interpretability while remaining robust to adversarial attacks.

Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation

Lucile Ter-Minassian, Liran Szlak, Ehud Karavani, Chris Holmes, Yishai Shimoni.
International Conference on Artificial Intelligence and Statistics Conference (AISTATS), 2025.

In Chapter 5, we address the need for interpretable and transparent methods in causal effect estimation from observational data, especially in policy-making contexts where high-stakes and the absence of ground truth labels demand greater accountability. We introduce

BICauseTree, a balancing method that identifies local natural experiments through an interpretable decision tree framework. By optimizing for covariate balance and reducing treatment allocation bias, BICauseTree can detect and exclude subgroups with positivity violations, offering a clear definition of the target population for valid inference and generalization.

Is Merging Worth It ? Securely Evaluating the InformationGain for Causal Dataset Acquisition

> Jake Fawkes*, **Lucile Ter-Minassian***, Desi Ivanova, Uri Shalit, Chris Holmes.
> *International Conference on Artificial Intelligence and Statistics Conference* (AISTATS), 2025.

In Chapter 6, we focus on optimizing data collection strategies to better answer causal questions. Specifically, we address the challenge of securely evaluating which datasets are most valuable for investigating the causal effect of a feature, without compromising sensitive information. We introduce a cryptographically secure, information-theoretic framework that quantifies the potential benefit of merging datasets in the context of heterogeneous treatment effect estimation. Our method thus allows data hosts to assess merge value without revealing raw data.

Other projects

Even though not included, I have been fortunate enough to have been part of several other works which I list below.

Ter-Minassian, L.*, Ghalebikesabi, S.*, Diaz-Ordaz, K., & Holmes, C. (2022) *Challenges and Opportunities of Shapley values in a Clinical Context*. ICML 2022 Workshop on Interpretable Machine Learning in Healthcare. Available at: `https://arxiv.org/abs/2306.14698`

Ter-Minassian, L., Ghalebikesabi, S., Diaz-Ordaz, K., & Holmes, C. (2022) *Explainable AI for survival analysis: a median-SHAP approach*. ICML 2022 Workshop on Interpretable Machine Learning in Healthcare. Available at: `https://arxiv.org/abs/2402.00072`

Pochinkov, N., Benoit, A., Agarwal, L., Majid, Z. A., & **Ter-Minassian, L.** (2024) *Extracting Paragraphs from LLM Token Activations*. NeurIPS 2024 MINT Workshop. arXiv preprint arXiv:2409.06328. Available at: `https://arxiv.org/abs/2409.06328`

Ter-Minassian, L. (Under review) Democratizing AI Governance: Balancing Expertise and Public Participation. Available at: https://arxiv.org/pdf/2502.08651

Chapter 2

Background

2.1 Scope of the thesis

Throughout this thesis, our focus will be on improving approximations of various truth concepts: causal truths, model truths, and true data distributions. These distinct notions of truth converge toward a shared goal—building AI systems that are more trustworthy and transparent. Causal inference seeks to reveal the true mechanisms behind observed data, allowing for robust, generalizable insights. Explainable AI, by contrast, opens up black-box models to human understanding, revealing how and why decisions are made. By addressing both *external* (real-world causal) and *internal* (model-based) forms of understanding, these approaches jointly tackle the epistemic challenges of safe, reliable AI. This background section begins by introducing key concepts in causal inference, transitions to fundamental notions in explainable AI, and concludes by explicitly connecting each paper to a specific type of truth approximation.
In this thesis, our work on explainability focuses on local explanation models, applied to both tabular and image data. In contrast, our contributions to causal inference are tailored specifically to tabular data, where structured covariates and treatment-outcome relationships can be explicitly modeled. This background section begins by introducing key concepts in causal inference, transitions to fundamental notions in explainable AI, and concludes by explicitly connecting each paper to a specific type of truth approximation.

2.2 Causal Inference Background

2.2.1 Fundamental Causal Concepts

Notation and Problem Formulation

Let X represent a set of observed covariates taking values in the space $\mathscr{X}$, T denote a binary treatment variable taking values in $\{0,1\}$, and Y represent the outcome of interest.

The outcome can be binary or continuous.

Potential Outcomes Framework

The Potential Outcomes Framework, often referred to as the Rubin Causal Model, conceptualizes causal effects by considering the outcomes that would manifest under different treatment conditions for each unit in a population [4]. For a binary treatment $T \in \{0,1\}$, the potential outcomes are denoted as $Y(1)$ and $Y(0)$, representing the outcomes if the unit receives the treatment ($T = 1$) or control ($T = 0$), respectively. The individual causal effect is defined as the difference between these potential outcomes $Y(1) - Y(0)$. However, since only one of these outcomes can be observed for each unit – a phenomenon known as the fundamental problem of causal inference – statistical methods are employed to estimate average causal effects across populations.

Directed Acyclic Graphs (DAGs)

DAGs serve as a cornerstone in Pearl's Structural Causal Model framework, providing a graphical representation of causal relationships among variables [3]. In a DAG, nodes represent variables, and directed edges indicate causal influences, with the acyclic property ensuring the absence of feedback loops. This structure facilitates the identification of causal pathways and the application of graphical criteria, such as d-separation.

d-Separation is a graphical criterion used in DAGs to determine whether a treatment variable T is conditionally independent of a set of potential confounders X, given an outcome Y. This concept is fundamental in causal inference and probabilistic graphical models, as it allows researchers to infer conditional independencies directly from the graph's structure without resorting to algebraic computations. In essence, d-separation helps identify whether information about T provides any additional insight into X once Y is known. If T and X are d-separated by Y, then T and X are conditionally independent given Y. This process is crucial for identifying confounding, assessing the validity of causal assumptions, and understanding the underlying causal relationships represented by the DAG.[3].

Causal Identifiability Conditions

Establishing causal relationships necessitates certain identifiability conditions to ensure that the estimated effects reflect true causal influences rather than spurious associations.

Consistency

The consistency assumption posits that the observed outcome for a unit under a particular treatment level corresponds exactly to the potential outcome under that treatment. Formally, if Y denotes the observed outcome and T the treatment assignment, then:

$$Y = Y(T).$$

This implies that the treatment, as implemented in the study, is identical to the treatment defined in the potential outcomes framework.

Positivity

Positivity, also known as the overlap condition, requires that every unit has a non-zero probability of receiving each treatment level, given their covariate profile. Mathematically, this is expressed as:

$$0 < P(T = t \mid X = x) < 1 \quad \forall t \in \{0,1\}, \forall x \in \mathscr{X},$$

where X represents the covariates. This condition ensures that comparisons between treatment groups are meaningful across all levels of covariates.

Exchangeability

Exchangeability, also known as the assumption of no unmeasured confounding, asserts that conditional on observed covariates X, the potential outcomes are independent of the treatment assignment:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X.$$

This implies that, within strata of X, the treatment assignment is as good as random, allowing for unbiased estimation of causal effects from observational data.
Formally, for any values x in the support of X, we have:

$$\mathbb{P}(Y(1), Y(0) \mid T = 1, X = x) = \mathbb{P}(Y(1), Y(0) \mid T = 0, X = x).$$

Example. Suppose we are interested in the effect of a job training program ($T = 1$) on employment status (Y). If all individuals are observed along with their education level X, and education is the only confounder, then conditional on education level, treatment assignment is assumed to be independent of potential employment outcomes. That is:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid \text{Education}.$$

In this case, comparing treated and untreated individuals within each education stratum yields an unbiased estimate of the causal effect.

Rubin and Pearl's Approaches: A Brief Comparison

The Rubin Causal Model (RCM) and Pearl's Structural Causal Model (SCM) offer distinct yet complementary frameworks for causal inference. RCM, grounded in the potential outcomes framework, emphasizes randomised experiments and the use of statistical methods to estimate causal effects, often relying on assumptions like ignorability and employing tools such as propensity scores [34]. The *propensity score* is defined as the probability of receiving treatment given observed covariates, $\pi(X) = \mathbb{P}(T = 1 \mid X)$, where T is the treatment indicator and X represents covariates.
In contrast, SCM utilizes Directed Acyclic Graphs (DAGs) to explicitly model causal relationships and dependencies among variables, providing a graphical criterion (d-separation) to assess conditional independencies and guide the identification of causal effects [3].

2.2.2 Causal Effect Estimation

Testing for assumptions

Before estimating causal effects, it is essential to test the above mentioned assumptions. Some assumptions can be empirically tested using observed data, while others are inherently untestable and must be justified through subject-matter expertise.
Positivity violations occur when certain subgroups rarely or never receive some treatments, leading to data sparsity. Such violations can be diagnosed by inspecting the distribution of treatment assignments across covariates [35].
The exchangeability assumption pertains to unobserved variables, it thus cannot be empirically tested and must be justified based on study design and domain expertise [5].

Evaluating Balance and Overlap

Effective causal inference requires evaluating whether balancing procedures successfully mitigated confounding. *Balance* refers to the similarity in the distribution of covariates between treatment groups after adjustment, ensuring that comparisons are not driven by systematic differences. *Overlap* directlty relates to the positivity assumption – it denotes the presence of comparable treated and control units across the covariate space, such that treatment effects can be reliably estimated.
Absolute Standardised Mean Differences—often displayed in “Love plots”—indicate acceptable balance [36]. Examining propensity score distributions ensures sufficient overlap between groups, as regions with poor common support yield unreliable causal estimates.

Addressing Positivity Violations

A common approach to adress positivity violations is to exclude subgroups where the modelled treatment assignment probabilities are near zero or one. This method is studied by *Crump et al.* and is called *propensity thresholding* [37]. However, since the exclusion criteria are derived from propensity score models, which are not always interpretable, it becomes challenging to delineate the specific subpopulations to which the findings can be generalised based on observable features. Furthermore, this exclusion reduces the effective sample size and may limit the generalizability of the findings. Therefore, careful consideration is needed to balance bias and variance when addressing positivity violations [38].

Average Treatment Effects (ATE)

The Average Treatment Effect (ATE) quantifies the expected difference in potential outcomes between the treated and control conditions across the entire population. Let $Y(1)$ and $Y(0)$ denote the potential outcomes under treatment and control, respectively. Let X be a vector of observed covariates, and P the joint distribution over these covariates in the population. Then the ATE is defined as:

$$\mathrm{ATE} = \mathbb{E}[Y(1) - Y(0)] = \int \mathbb{E}[Y(1) - Y(0) \mid X = x] \, dP(x),$$

where $dP(x)$ denotes integration with respect to the marginal distribution of covariates X. Under the assumptions of consistency, positivity, and exchangeability, the ATE can be identified from observational data using methods such as regression adjustment, inverse probability weighting, or matching.

Conditional Average Treatment Effects (CATE)

The Conditional Average Treatment Effect (CATE) characterizes how the treatment effect varies across subpopulations defined by covariates. For a specific covariate value x, the CATE is defined as:

$$\mathrm{CATE}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x],$$

where X is the vector of observed covariates. Unlike the ATE, which averages over the population, the CATE focuses on treatment effects conditional on $X = x$, capturing heterogeneity in causal effects. Estimating the CATE typically involves modeling interactions between the treatment and covariates, using flexible methods such as tree-based models or nonparametric regression.

2.2.3 Methods for Causal Inference from Observational Data

Matching and Propensity Score Methods

Matching methods create comparable treatment and control groups by pairing units with similar covariate profiles, effectively emulating randomised controlled trials. Propensity score matching, introduced by Rosenbaum and Rubin [34], estimates treatment assignment probability conditional on observed covariates to reduce confounding bias. Beyond matching, propensity scores enable various adjustment techniques including stratification, weighting, and covariate adjustment in regression models. These methods assume all confounders are measured and correctly specified.

2.2.4 Covariate Balancing Techniques

A central challenge in estimating causal effects from observational data is addressing confounding due to systematic differences in covariate distributions between treated and control units. Covariate balancing techniques aim to mitigate this bias by constructing weighted or matched samples in which the distribution of covariates is similar across treatment groups.

One widely used approach is *weighting*, particularly through the method of **Inverse Probability of Treatment Weighting** (IPTW). The core idea is to reweight the observed data to create a *synthetic pseudo-population* in which the treatment assignment is independent of the covariates, as it would be under randomized assignment.

Let $T_i \in \{0,1\}$ denote the treatment indicator for unit i, and let X_i be the associated vector of observed covariates. The propensity score is defined as the conditional probability of receiving treatment given covariates:

$$\pi(X_i) = \mathbb{P}(T_i = 1 \mid X_i).$$

Each unit is assigned a weight:

$$w_i = \frac{T_i}{\pi(X_i)} + \frac{1 - T_i}{1 - \pi(X_i)},$$

so that treated units ($T_i = 1$) are weighted by $1/\pi(X_i)$, and control units ($T_i = 0$) by $1/(1 - \pi(X_i))$. These weights adjust for the varying probabilities of treatment across levels of X, upweighting units that are underrepresented in their respective treatment group.

The resulting weighted sample can be interpreted as a pseudo-population in which, under the assumption of exchangeability (i.e., no unmeasured confounding), treatment is independent of covariates. In this population, the average treatment effect (ATE) can be consistently estimated as:

$$\widehat{\text{ATE}}_{\text{IPTW}} = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot Y_i^{\text{obs}} \cdot (2T_i - 1),$$

where Y_i^{obs} is the observed outcome and $(2T_i - 1) \in \{-1, 1\}$ indicates the treatment contrast.

While IPTW can be effective, it is sensitive to extreme values of the propensity score $\pi(X_i)$, which can result in large weights and high variance. This has motivated the development of stabilized weights and alternative balancing methods that directly target covariate balance without relying solely on propensity score estimation [39].

2.2.5 Modern Machine Learning Approaches

Causal Trees and Forests

Causal Trees extend traditional decision tree methods—specifically the CART (Classification and Regression Trees) framework—to the estimation of heterogeneous treatment effects. Whereas standard CART aims to partition the covariate space X to minimize prediction error of the observed outcome Y, Causal Trees seek to identify splits that expose variation in the Conditional Average Treatment Effect (CATE), defined as

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

The splitting criterion is adapted to maximize heterogeneity in estimated treatment effects between child nodes rather than minimize outcome variance alone [40].

Formally, Causal Trees rely on an *honest* estimation procedure: the data is partitioned into a *splitting set* and an *estimation set*. The splitting set is used to determine the tree structure, while treatment effects in each leaf are estimated using the estimation set to avoid overfitting. A candidate split is chosen to maximize the expected reduction in the mean squared error (MSE) of the CATE estimator, evaluated on the estimation sample.

Causal Forests generalize this framework by averaging predictions from a large ensemble of causal trees, each built on a bootstrap sample and incorporating random feature selection at each split. The resulting forest provides a consistent and asymptotically normal estimator of the CATE function [41]. Unlike standard Random Forests, Causal Forests retain the honesty principle: each tree is grown with a subsample for splitting and a separate subsample for estimation, improving statistical validity and allowing for valid inference.

Meta-Learners

Meta-learners are strategies that leverage machine learning models to estimate treatment effects, particularly useful for assessing how treatment impacts outcomes across different subgroups [42].

T-Learner This approach fits separate models for the treatment and control groups. Specifically, it estimates two models: one for the treated group $Y(1)$ and one for the control group $Y(0)$. The treatment effect is then computed as the difference between these two predicted outcomes:

$$\hat{\tau}_T(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X),$$

where $\hat{\mu}_1(X) = \hat{\mathbb{E}}[Y \mid T = 1, X]$ and $\hat{\mu}_0(X) = \hat{\mathbb{E}}[Y \mid T = 0, X]$.

S-Learner This method incorporates the treatment indicator T as a feature within a single model, estimating the outcome as a function of both covariates and treatment:

$$\hat{\mu}(X, T) = \hat{\mathbb{E}}[Y \mid X, T].$$

The treatment effect is then estimated by computing the difference in predicted outcomes when $T = 1$ and $T = 0$:

$$\hat{\tau}_S(X) = \hat{\mu}(X, 1) - \hat{\mu}(X, 0).$$

X-Learner This method is particularly beneficial when there are imbalances between the treated and control groups. It proceeds in several steps:

1. Estimate outcome models $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ using the treated and control groups, respectively.

2. Compute imputed treatment effects for each group:

   - For treated units: $D^1 = Y - \hat{\mu}_0(X)$.
   - For control units: $D^0 = \hat{\mu}_1(X) - Y$.

3. Fit models to predict D^1 and D^0 as functions of X, yielding $\hat{\tau}_1(X)$ and $\hat{\tau}_0(X)$, respectively.

4. Combine these estimates using a weighting function $g(X)$, often based on the estimated propensity score $\hat{\pi}(X)$:

$$\hat{\tau}_X(X) = g(X) \cdot \hat{\tau}_0(X) + (1 - g(X)) \cdot \hat{\tau}_1(X).$$

This approach allows the X-Learner to address potential bias due to group size imbalances and more accurately estimate heterogeneous treatment effects across individuals.

Doubly Robust Estimators

Doubly Robust (DR) estimators combine outcome modeling with propensity score weighting, providing consistent estimates of the average treatment effect (ATE) if either model is correctly specified [43].
Let:

- Y: observed outcome,
- $T \in \{0,1\}$: treatment indicator,
- X: covariates,
- $\pi(X) = P(T = 1 \mid X)$: propensity score,
- $\mu_t(X) = \mathbb{E}[Y \mid T = t, X]$: outcome regression for $t \in \{0,1\}$.

The doubly robust estimator for the ATE is:

$$\hat{\tau}_{DR} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)(1 - \hat{\pi}(X_i))} \cdot (Y_i - \hat{\mu}_{T_i}(X_i)) \right].$$

This estimator remains consistent if either the propensity score model $\hat{\pi}(X)$ or the outcome regression models $\hat{\mu}_t(X)$ are correctly specified.

Targeted Maximum Likelihood Estimation (TMLE)

TMLE is a semi-parametric method that integrates machine learning algorithms with a targeted update step to optimize causal parameter estimation, reducing bias and variance [44, 45].
Let:

- Y: observed outcome,
- $T \in \{0,1\}$: treatment indicator,

- X: covariates,
- $\pi(X) = P(T = 1 \mid X)$: propensity score,
- $\mu_t(X) = \mathbb{E}[Y \mid T = t, X]$: outcome regression for $t \in \{0, 1\}$.

The TMLE procedure involves:

1. Estimating initial outcome regressions $\hat{\mu}_t(X)$ and propensity score $\hat{\pi}(X)$.
2. Defining the clever covariate:
$$H(T,X) = \frac{T}{\hat{\pi}(X)} - \frac{1-T}{1-\hat{\pi}(X)}.$$
3. Updating the initial outcome model via a logistic regression of Y on $H(T,X)$ with offset $\text{logit}(\hat{\mu}(T,X))$, yielding an updated estimate $\tilde{\mu}(T,X)$.
4. Computing the TMLE estimate of the ATE:
$$\hat{\tau}_{TMLE} = \frac{1}{n} \sum_{i=1}^{n} \left[\tilde{\mu}(1,X_i) - \tilde{\mu}(0,X_i)\right].$$

TMLE is consistent if either the outcome model $\hat{\mu}_t(X)$ or the propensity score model $\hat{\pi}(X)$ is correctly specified, and it achieves the semiparametric efficiency bound when both are correctly specified.

2.2.6 Challenges of Causal Inference

Transparency and Interpretability A fundamental challenge in causal inference is navigating the trade-off between model complexity and interpretability. Some methods, such as matching or propensity score weighting [46], are relatively transparent and offer intuitive insights into how treatment effects are estimated. However, these simpler approaches may struggle to capture the complex, nonlinear relationships present in high-dimensional observational data. In contrast, modern machine learning methods—like causal forests [40] or deep learning—can flexibly model such complexities, often yielding more accurate effect estimates. Yet, their inner workings are typically opaque, making it difficult to understand how or why a given causal conclusion is reached.

This tension is particularly problematic in high-stakes domains such as healthcare, policy-making, and social interventions, where stakeholders require more than just accurate numbers—they need to trust that these numbers reflect real-world mechanisms. In such settings, interpretability is not merely a convenience but an ethical necessity. As Rudin argues in her call to avoid black-box models in critical applications [28], and as Pearl emphasizes in The Book of Why [2], causal inference must strive to be both accurate and transparent. This thesis explores methods that seek to balance this trade-off, aiming to improve estimation performance without sacrificing the interpretability needed for trustworthy decision-making.

Granularity of Causal Effects A key limitation in causal inference is the reliance on aggregate measures like Average Treatment Effect (ATE) or Conditional Average Treatment Effect (CATE), which provide only summary impacts of a treatment variable [47]. These measures fail to capture the complex pathways through which treatments influence outcomes. Interventions often affect outcomes through multiple simultaneous mechanisms—some beneficial, others harmful—which aggregate measures obscure. For example, medications may improve clinical outcomes through intended pathways while causing adverse effects through others. The inability to decompose effects into constituent pathways represents a significant limitation, particularly when specific mechanisms have scientific or policy importance [48]. This coarse granularity masks how treatments interact with mediating variables, which can vary substantially across individuals, thereby limiting insights into causal structures beyond overall importance measures for the treatment variable.

Positivity Violations and Generalizability A critical challenge in causal inference is the failure to explicitly characterize the target population for which causal effects are identifiable. This issue stems from violations of the positivity assumption—that all units have a non-zero probability of receiving each treatment level across all covariate values [49]. In practice, some subpopulations experience deterministic or near-deterministic treatment assignment, creating regions in the covariate space where counterfactual outcomes cannot be reliably estimated. Even when positivity holds technically, practical violations occur when treatment probabilities are extremely small for certain subgroups, leading to unstable estimates. These violations directly limit the generalizability of causal findings, as effect estimates may only apply to specific subpopulations where sufficient overlap exists. When findings are inappropriately extrapolated to populations that differ from the study sample, misleading conclusions often result [50]. The dual challenge remains: identifying regions where causal effects can be reliably estimated and precisely defining the population to which these estimates generalize.

Privacy Constraints and Data Integration The distributed nature of data across institutions poses significant challenges for causal inference. Critical information for adjusting confounding often resides in separate data systems subject to privacy regulations and proprietary restrictions [51]. This fragmentation limits sample sizes and potentially introduces selection bias when analyses are restricted to single-institution data. Even when technical solutions for data sharing exist, legal and ethical constraints may prevent the pooling of individual-level data, particularly in domains like healthcare where privacy concerns are paramount. Traditional anonymization techniques often prove insufficient, as they either compromise privacy through re-identification risks or degrade data utility through excessive noise. Moreover, the prospective value of merging datasets remains difficult to quantify without revealing sensitive information, creating a circular problem where data cannot be shared without knowing its value, yet its value cannot be determined without sharing. These constraints systematically limit the evidence base for causal questions, particularly those requiring large, diverse samples to detect heterogeneous effects or rare outcomes [52].

Epistemic Uncertainty Causal inference faces a fundamental epistemic uncertainty absent in predictive tasks: the impossibility of observing counterfactual outcomes [7]. This unverifiability introduces uncertainty that standard statistical approaches inadequately capture. Multiple interacting sources—sampling variability, model misspecification, unobserved confounding, and measurement error—create complex uncertainty patterns that conventional confidence intervals fail to represent. Limited overlap between treatment groups further compounds this problem, as extrapolation beyond common support regions relies on untestable assumptions. The challenge extends to effectively communicating uncertainty to decision-makers, particularly in high-dimensional settings or sequential decision problems [53]. This fundamental inability to validate causal claims against ground truth necessitates careful consideration of identification assumptions and sensitivity analyses when making causal inferences.

2.3 Explainable Artificial Intelligence Background

2.3.1 An Overview of the Families of XAI Methods

Selecting an appropriate XAI method depends on the model's characteristics and the specific interpretability objectives. Determining the purpose of the explanation is crucial:

- Model decisions: Elucidating the model's outputs or predictions.
- Internal mechanisms: Understanding the internal workings or representations within the model.

Interpretable Models

Inherently interpretable models, such as linear regression or decision trees, offer transparent relationships between inputs and outputs, facilitating straightforward explanations [54]. The main challenge of interpretable models remain the potential deprecation of performance.

Post-Hoc Explanation Methods

For complex, less interpretable models, post-hoc methods provide insights into model behaviour. These methods are categorised along several dimensions:

- *Local vs. Global Interpretability*: Local methods explain individual predictions, while global methods characterize the model's overall behaviour.
- *Model-Agnostic vs. Model-Specific*: Model-agnostic methods can explain any black-box model by treating it as a function that maps inputs to outputs. Model-specific methods leverage the internal structure of particular architectures. For instance, various gradient-based methods are specific to neural networks: [55, 56].

In this thesis, we focus on *local, model-agnostic* XAI methods. These methods can output different types of explanations, including but not limited to:

- *Local model approximations*: A simple surrogate model, $g(x)$, is fitted to approximate the black-box model $f(x)$ in a neighbourhood around the prediction point x. For example, LIME [22] fits a sparse linear model around an image or tabular instance by perturbing its features and weighting the perturbed samples by their proximity to x.
- *Additive feature attribution methods*: A model-agnostic estimator describes how the model outcome at a point x changes when some of its features are removed or replaced with values sampled from a reference distribution $n(\tilde{x} \mid x)$. For instance, SHAP [57] computes feature attributions by estimating each feature's marginal contribution to the prediction, using ideas from cooperative game theory. On a loan application, SHAP might assign 70 points of a 600 credit score to high income and subtract 30 due to recent delinquencies, yielding an additive decomposition of the model output.
- *Concept-based attribution methods*: TCAV quantifies how sensitive a model's prediction for a class k is to a human-defined concept C [58]. For example, one may ask: how much does the concept "striped" influence the model's prediction of the class "zebra"?

Given a set of positive examples P_C representing concept C, and negative examples N_C, activations are extracted from a chosen layer l:

$$A_P = \{f_l(x) \mid x \in P_C\}, \quad A_N = \{f_l(x) \mid x \in N_C\},$$

where $f_l(x)$ denotes the activation of input x at layer l. A linear classifier is trained to distinguish between A_P and A_N, yielding a hyperplane with normal vector v_C^l, the Concept Activation Vector (CAV), representing the direction of concept C in the activation space.

For an input x and class k, the directional derivative of the logit $h_{l,k}(f_l(x))$ along v_C^l measures sensitivity to concept C:

$$S_{C,k,l}(x) = \nabla h_{l,k}(f_l(x)) \cdot v_C^l.$$

The CAV importance score aggregates this sensitivity over a set $\mathscr{X}_k$ of inputs belonging to class k:

$$\mathrm{TCAV}_{C,k,l} = \frac{1}{|\mathscr{X}_k|} \sum_{x \in \mathscr{X}_k} \mathbb{I}[S_{C,k,l}(x) > 0],$$

where $\mathbb{I}[\cdot]$ is the indicator function. A higher score indicates greater influence of concept C on the model's prediction for class k.

This thesis will particularly focus on *local model approximations* and *additive feature attribution methods*. Let $x = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ denote the feature vector of the test instance, where x_j refers to the value of the j-th feature. Throughout this thesis, we distinguish between the individual components x_j of a single test point and the set $\{\mathbf{x}_i\}_{i=1}^N$, which denotes a collection of covariate vectors (i.e., the background or reference dataset). The model of interest is a black-box function $f : \mathbb{R}^m \to \mathbb{R}$ whose predictions we aim to explain. To do so, a local surrogate model $g : \mathbb{R}^m \to \mathbb{R}$ is fitted around the point x using perturbations sampled from a neighborhood distribution $n(x^* \mid x)$. In local approximation methods such as LIME [59], g is typically chosen to be linear, i.e., $g(x) = \sum_{j=1}^m \beta_j x_j$, and the coefficients β_j are obtained by minimizing the squared loss $\mathbb{E}_{n(X^* \mid x)}[(g(X^*) - f(X^*))^2] \approx \sum_{i=1}^L (g(x_i^*) - f(x_i^*))^2$.

By contrast, additive feature attribution methods, such as SHAP [57], define the explanation model over a simplified binary input space $\{0,1\}^m$, where $z_j = 1$ indicates the presence of feature j and $z_j = 0$ its absence. The surrogate model takes the form

$$g(z) = \phi_0 + \sum_{j=1}^m \phi_j z_j,$$

where ϕ_j quantifies the marginal contribution of feature j to the prediction.

These coefficients are computed as Shapley values from cooperative game theory, which fairly distribute the model output among input features. Formally, for a value function $v(T,x)$ representing the expected model output when a subset T of features is known, the Shapley value for feature j is:

$$\phi_v(j,x) = \sum_{T \subseteq \{1,\dots,m\} \setminus \{j\}} \frac{|T|!\,(m-|T|-1)!}{m!} \left[v(T \cup \{j\}, x) - v(T,x)\right].$$

The value function is typically instantiated as the expected output of the model when missing features are marginalized out using a reference distribution:

$$v(T,x) = \mathbb{E}_{r(X^*_{\overline{S}} \mid x_S)} \left[f(x_T, X^*_{\overline{S}})\right],$$

where $\overline{S} = \{1,\dots,m\} \setminus T$ and $r(X^*_{\overline{S}} \mid x_S)$ denotes either the marginal or conditional distribution of the missing features.

While both LIME and SHAP approximate the black-box model locally using linear models, they differ in the representation of the input: LIME operates directly in the original feature space, whereas SHAP interprets inputs in terms of feature inclusion or exclusion relative to a background distribution. Further details are provided in Chapter 3.

Challenges of Post-Hoc Explainability

Post-hoc explainability methods for black-box models face several fundamental challenges that impact their reliability and utility. We focus on two key challenges: fidelity to the model versus fidelity to the data, and the limitations of locality in explanations.

Model Fidelity versus Data Fidelity A critical challenge in post-hoc XAI is the tension between explanations that are "true to the model" versus those that are "true to the data" [60, 61]. An explanation is said to be true to the model if it accurately reflects the model's internal decision logic—regardless of whether that logic aligns with real-world patterns or human intuition. In contrast, an explanation is true to the data if it captures structure that genuinely exists in the underlying data distribution, even if the model itself fails to do so. This distinction is especially important in counterfactual or perturbation-based explainability methods. For instance, consider a clinical risk prediction model trained on real-world health data. If we generate explanations by randomly perturbing input features—say, by sampling smoking status independently from its marginal distribution—we may evaluate the model on implausible inputs, such as a *smoking 2-year-old patient*. This point lies outside the support of the training data, and any attribution made there may be faithful to the model's behavior but meaningless or misleading from a data (and clinical) perspective. Explanations in such regions may highlight artifacts in the model rather than insights grounded in reality.

A fundamental property that model-faithful explanations should satisfy is the "dummy property" [62], which stipulates that a feature that has no influence on the model's prediction should receive zero attribution, regardless of its statistical relationship with other features in the data. Formally, if $\frac{\partial f(x)}{\partial x_j} = 0$ for all inputs in a neighborhood of x, then $\phi_j(x) = 0$. However, this property is often violated in practice for several reasons:

- *Approximation error*: Local surrogate models (e.g., LIME) may incorrectly attribute importance to irrelevant features due to the inherent error in approximating complex models with simpler ones.
- *Reference data sampling*: When simulating feature absence by sampling from reference distributions, correlations in the training data may cause attributions to be assigned to features that the model does not actually use. For instance, if features x_i and x_j are highly correlated in the training data, removing x_i may alter the prediction not due to the model's direct reliance on x_i, but because x_j tends to take values that are conditionally dependent on those of x_i—and this dependence is broken upon removal.
- *Off-manifold predictions*: Imputation strategies may generate synthetic data instances that lie far from the data manifold, causing explanations to be based on model predictions for implausible inputs. Since most machine learning models are not constrained to behave reasonably on inputs that lie outside the training distribution, these off-manifold predictions may introduce significant distortions into the resulting explanations.

These issues highlight the fundamental difficulty of disentangling model behaviour from data distribution characteristics in post-hoc explanations.

Limitations of Locality The notion of "locality" in explanation methods presents another significant challenge. Truly local explanations aim to describe model behaviour in the immediate vicinity of a specific instance x. However, this goal often conflicts with the mechanisms used to construct explanations:

- *Neighborhood definition*: The definition of the "local neighborhood" around an instance is often arbitrary and can significantly impact the resulting explanation.

Different choices of neighborhood size or sampling strategy may lead to substantially different explanations for the same prediction.

- *Feature perturbation conflicts*: Methods that rely on perturbing features to observe changes in model output may explore regions of the input space that are far from the local neighborhood of interest. This is particularly problematic when features are correlated, as perturbing one feature while holding others constant may create synthetic instances that are highly improbable or impossible in the real data distribution.

- *Manifold inconsistency*: Real-world data often lies on or near a lower-dimensional manifold within the higher-dimensional feature space. Local explanation methods that generate perturbations without respecting this manifold structure may base their explanations on model behaviour at points that are technically "nearby" in the feature space but represent implausible or even meaningless instances.

These locality challenges are further compounded when explanation methods employ imputation mechanisms to simulate feature absence, as the imputed values may draw the perturbed instances away from the region of interest, undermining the local nature of the explanation. Addressing these challenges requires careful consideration of the specific objectives of the explanation and the trade-offs between model fidelity, data fidelity, and locality. Methods that explicitly account for data manifold structure and model behaviour simultaneously, such as counterfactual explanations with plausibility constraints [63] or manifold-aware feature attributions [64], offer promising directions for more reliable post-hoc explainability.

2.4 Evaluation of our Frameworks

Evaluating the reliability and utility of methods in explainable artificial intelligence (XAI) and causal inference involves distinct methodological challenges, especially in the absence of ground truth. Unlike predictive tasks, where accuracy can be directly measured, these domains prioritise interpretability, robustness, and alignment with domain knowledge—qualities that are inherently harder to quantify.

In XAI, the absence of standardised benchmarks makes systematic comparison difficult. Explanations may aim for fidelity to the model or plausibility with respect to the data distribution, but these goals often conflict. Additionally, black-box models may behave unpredictably off-manifold, producing unstable or misleading explanations.

Causal inference faces its own epistemic constraint: counterfactuals cannot be observed. As a result, evaluation must rely on assumptions, benchmarks, or indirect diagnostics such as covariate balance and sensitivity analyses. Further complications arise from reliance on untestable assumptions like unconfoundedness and positivity, and from the context-specific nature of causal estimands.

To address these limitations, we employ a multifaceted evaluation strategy. For XAI, we combine fidelity metrics, adversarial robustness checks, and OOD detection, alongside qualitative assessment. For causal inference, we assess covariate balance, support overlap, and consistency with known benchmarks or alternative estimators. In both cases, sensitivity analyses are used to explore the dependence of results on key assumptions and hyperparameters.

This strategy aims to establish empirical and theoretical grounds for trust in our methods, despite the fundamental limitations posed by the lack of observable ground truth.

2.5 Thesis Structure and Contributions

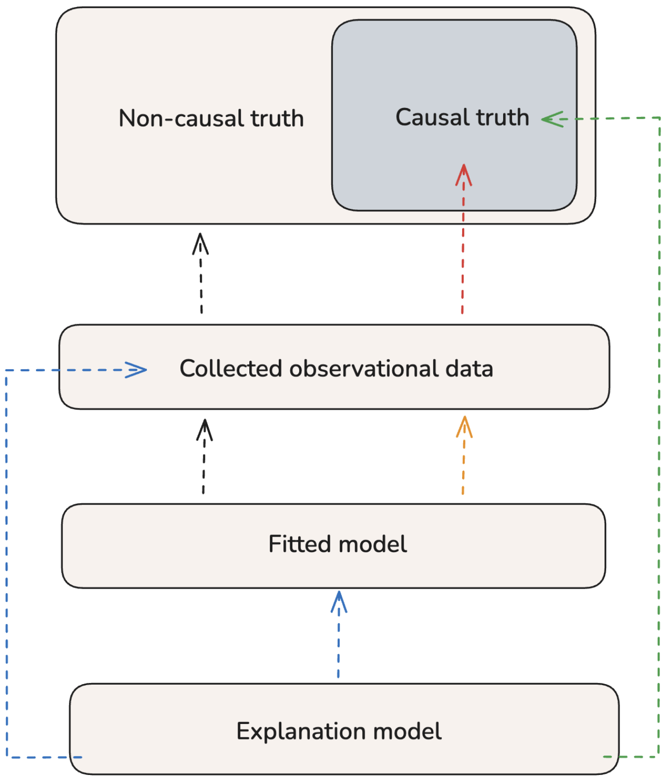


Figure 2.1: Overview of the themes approached in this thesis and their inter-connexions.

This thesis examines the challenge of approximating different types of truth in machine learning and causal inference. As illustrated in Figure 2.1, we identify three distinct but interconnected notions of truth: model truth, data-based truth, and causal truth. Each paper in this thesis addresses a specific aspect of these truth approximations, collectively advancing our understanding of explainability, interpretability, and causal inference.

Our journey through these different facets of truth follows a progression from model explainability to causal understanding. We begin by examining post-hoc explanations of black-box models, progress to causally explaining fitted models, advance to developing transparent causal inference methods, and conclude with techniques for optimizing data collection to better approximate causal truths.

Chapter 3 represented by the blue pathway in Figure 2.1 investigates the fundamental tension between explanation models that are faithful to the underlying data versus those faithful to the black-box model being explained. By introducing Neighbourhood SHAP values, we address limitations in conventional Shapley analysis regarding locality and on-manifold explainability. Our contribution resolves a critical gap in existing XAI methods by providing more meaningful sparse feature attributions that better reflect local model behaviour and demonstrate increased robustness against adversarial classifiers. This work fundamentally questions what it means for an explanation to be truly local.

Chapter 4 illustrated by the green pathway bridges the gap between predictive modelling and causal understanding. Through our Path-Wise Shapley effects PWSHAP framework, we extend model explainability to include causal pathways by augmenting predictive models with user-defined directed acyclic graphs. This novel approach enables practitioners to gain causal insights from complex black-box models, particularly valuable in sensitive domains where specific variables such as treatment effects or protected attributes require special scrutiny. By establishing error bounds and demonstrating applications in local bias and mediation analyses, this work advances our ability to extract causal understanding from predictive models.

Chapter 5, depicted in the orange pathway, addresses the challenge of transparent causal effect estimation directly from observational data. The BICauseTree method we introduce represents a significant advancement in interpretable causal inference by identifying clusters where natural experiments occur locally. Unlike post-hoc approaches to black-box models, our decision tree-based method offers inherent interpretability while improving covariate balance and reducing treatment allocation bias. A key contribution is the method's ability to detect subgroups with positivity violations and provide explicit covariate-based definitions of the target population, enhancing both transparency and generalizability of causal inferences.

Chapter 6, shown in the red pathway, examines how we can better approximate causal truth through strategic data integration. Recognizing that merging datasets across institutions presents privacy challenges yet offers potential benefits for causal estimation, we develop a cryptographically secure approach for quantifying the prospective value of data merges. Our Expected Information Gain framework, compatible with differential privacy, represents the first privacy-preserving method specifically designed for dataset acquisition in causal estimation contexts. This work addresses the critical challenge of reducing epistemic uncertainty and improving treatment overlap without compromising sensitive information.

Together, these papers form a coherent exploration of how we can better approximate various forms of truth in machine learning and causal inference. From improving model explanations to enhancing causal understanding, from developing transparent models to optimizing data collection, this thesis contributes methodological innovations that address key limitations in current approaches while advancing both theoretical understanding and practical applications.

Chapter 3

On Locality of Local Explanation Models

Abstract

Shapley values provide model agnostic feature attributions for a model outcome at a particular instance by simulating feature absence under a global population distribution. The use of a global population can lead to potentially misleading results when local model behaviour is of interest. Hence we consider the formulation of neighbourhood reference distributions that improve the local interpretability of Shapley values.

By doing so, we find that the Nadaraya-Watson estimator, a well-studied kernel regressor, can be expressed as a self-normalised importance sampling estimator. Empirically, we observe that Neighbourhood SHAP values identify meaningful sparse feature relevance attributions that provide insight into local model behaviour, complimenting conventional Shapley analysis. They also increase on-manifold explainability and robustness to the construction of adversarial classifiers.

3.1 Introduction

The ability to correctly interpret a prediction model is increasingly important as we move to widespread adoption of machine learning methods, in particular within safety critical domains such as health care [12, 65]. In this paper, we consider the task of attributing the features $\{1,\ldots,m\}$ of a complex machine learning model $f:\mathbb{R}^m\to\mathbb{R}^l$, abstracted as a function that predicts a response given a test instance $x\in\mathbb{R}^m$, given only black-box access to the model. We especially focus on two popular model-agnostic feature removal based local explanation models, namely LIME [22] and SHAP [57]. However our findings are applicable to other local explanation models that we do not consider in this paper. As these methods are often described as fitting a local surrogate model to the black-box [66], a natural question is: how 'local' are local explanation methods?

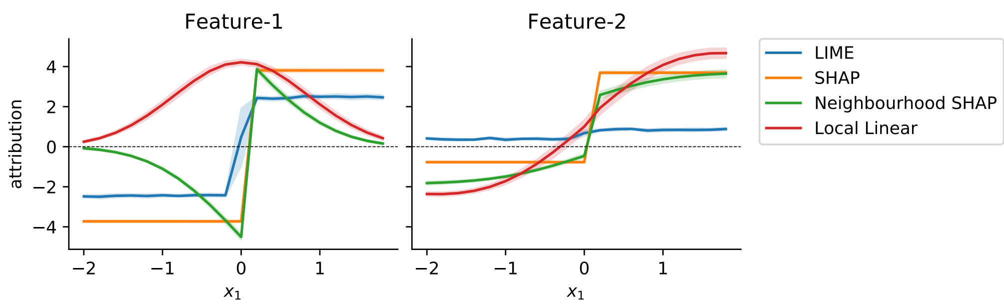


Figure 3.1: Attributions at $x=(x_1,2)$ with x_1 varying for a reference distribution of $X\sim\text{Normal}(0,1)$ and black-box $f(x)=\mathbb{I}(x_1>0)2x_2^2-\mathbb{I}(x_1\leq 0)x_2^2$. averaged over 10 runs displayed with 95% confidence intervals (see next section for details). While (Tabular) LIME and SHAP assign the same absolute attribution to Feature-1 no matter how large x_1 is, our neighbourhood approach takes its distance to the decision boundary into consideration. A local linear approximation to the black-box trained with a Ridge Regressor gives misleading attributions to Feature-1 for $-0.4<x_1<0$.

As a simple motivating example as to why this question matters, consider a black-box model given by $f(x)=\mathbb{I}(x_1>0)2x_2^2-\mathbb{I}(x_1\leq 0)x_2^2$ where $\mathbb{I}(\cdot)$ denotes the indicator function. When attributing the local feature importance at a test instance $x=(x_1,2)$, with x_2 fixed at 2, we would expect Feature-1 to receive a higher absolute attribution when x is closer to the decision boundary at $x_1=0$.

In Figure 3.1 we report the results on this example from LIME and SHAP as well as for our proposed 'Neighbourhood SHAP' approach. We observe that Neighbourhood SHAP assigns Feature-1 a smaller attribution, the higher the absolute value of x_1 is. SHAP and LIME, however, assign Feature-1 an attribution which is constant either side of $x_1=0$ which illustrates that these methods capture global model behaviour. The figure also shows that training a local linear approximation to the black-box [67, 68] is misleading since Feature-2 receives a significantly positive attribution for $x_1 \in [-0.4, 0]$, even though Feature-2 contributes clearly negatively to the model outcome whenever $x_1 < 0$.

This motivates the following contributions:

- We propose Neighbourhood SHAP (Section 3.3) which considers local reference populations for prediction points as a complimentary approach to SHAP. By doing so, we show that the Nadaraya-Watson estimator at x can be interpreted as an importance sampling estimator where the expectation is taken over the proposed neighbourhood. Empirically, we find that greater locality increases the number of model evaluations on the data manifold and with this the robustness of the attributions against adversarial attacks.
- We consider how smoothing can also be used to stabilise SHAP values (Section 3.4). We quantify the loss in information incurred by our smoothing procedure and characterise its Lipschitz continuity.

We refer the reader to Supplements A.2 for a more in-depth introduction to Shapley values.

3.2 Background

We begin with a short introduction to Shapley values – the quantity of interest of the SHAP optimisation procedure. For a pre-defined value function $v(T,x)$ that takes a set of features $T \subseteq \{1,...,m\}$ as input, the Shapley value $\phi_v(j,x)$ of feature j measures the expected change in the value function from including feature j into a random subset of features $S \subseteq \{1,...,m\} \setminus \{j\}$ (without j)

$$\phi_v(j,x) = \mathop{\mathbb{E}}_{p(S)} \left[v(S \cup \{j\}, x) - v(S,x) \right]$$

where the expectation is taken over the feature coalitions whose distribution is defined by

$$P(S) = \frac{|S|!\,(m-|S|-1)!}{m!}.$$

This probability distribution gives equal weight to coalitions of different sizes l,

$$P(\{S \mid |S| = k\}) = P(\{S \mid |S| = l\})$$

for $k,l \in \{0,...,m-1\}$.

The choice of value function for explanation-based modelling of feature attributions at an instance x has been the subject of recent debates [69, 70, 71]. The consensus is to take the expectation of the black-box algorithm f at observation x over the not-included features $\overline{S}$ using a reference distribution $r(X^*_{\overline{S}} \mid x_S)$ such that

$$v(S,x) = \underset{r(X^*_{\overline{S}} \mid x_S)}{\mathbb{E}} [f(x_S, X^*_{\overline{S}})]$$

for $\overline{S} := \{1,\ldots,m\}/S$ and the operation $(x_S, x_{\overline{S}})$ denoting the concatenation of its two arguments. Marginal Shapley values [57, 70] define $r(X^*_{\overline{S}} \mid x_S) := p(X^*_{\overline{S}})$ where p denotes the marginal data distribution. Conditional Shapley values [69] set the reference distribution equal to the conditional distribution given x_S, $r(X^*_{\overline{S}} \mid x_S) := p(X^*_{\overline{S}} | X^*_S = x_S)$.

All in all, the Shapley value $\phi(j,x)$ is characterised by the expected change in model output, comparing the output when we include j in the model, i.e. integrate out some randomly sampled features $\overline{S} \setminus \{j\}$, with the model output where feature j is not included, i.e. we integrated out some randomly sampled features including j, $\overline{S}$ $\phi(j,x) \in \mathbb{R}$ and is defined as:

$$\phi(j,x) = \underset{p(S)}{\mathbb{E}} \left[\underset{r(X^*_{\overline{S}\setminus\{j\}} \mid x)}{\mathbb{E}} [f(x_{S\cup\{j\}}, X^*_{\overline{S}\setminus\{j\}})] - \underset{r(X^*_{\overline{S}} \mid x_S)}{\mathbb{E}} [f(x_S, X^*_{\overline{S}})] \right]$$

As we see, Shapley values are computed by estimating the change in model outcome when some features are integrated out over the reference distribution $r(X^*_{\overline{S}} \mid x_S)$, which has so far been defined as either the marginal or conditional *global* population. For marginal Shapley values, the interpretation simplifies: The Shapley value of feature j is the expected change in model outcome when we sample a random individual x^* from the global statistical population and set its feature j equal to x_j (after we already set a random set of features $S \in \{1,\ldots,m\} \setminus \{j\}$ equal to x_S). This motivates our main idea–which we introduce in Section 3.3–of neighbourhood distributions where we instead sample a random individual from the immediate neighbourhood of x, as outlined in the next section.

Computing Shapely values is challenging in high-dimensional feature spaces, which motivates the widely adopted KernelSHAP approach [57] that estimates the Shapley values of all features by empirical risk minimisation of

$$\underset{p(S)}{\mathbb{E}} [(\underset{r(X^*_{\overline{S}} \mid x_S)}{\mathbb{E}} [f(x_S, X^*_{\overline{S}})] - g(S))^2] \approx \sum_{l=1}^{L} \sum_{i=1}^{C} w_i (f(x_{S_i}, x^*_{l,\overline{S}_i}) - g(S_i))^2 \tag{3.1}$$

where $g(S) = \phi_0 + \sum_{j=1}^m \phi(j,x) \mathbb{I}(j \in S)$ is a linear explanation model with the Shapley values as its coefficients, $\{x^*_{l,\overline{S}_i}\}_{l=1}^L$ are i.i.d. L draws from the respective global reference distributions, $\{S_i\}_{i=1}^{C}$ is a set of size C of sampled coalitions, and the weights w_i are defined by the KernelSHAP weights [57].

LIME optimises a similar generalised expectation – also sampling references from a global distribution. To improve local fidelity of Tabular LIME, [22] propose to define the weights as $w_i = \exp(-(|S_i| - m)^2/\sigma^2)$ for a bandwidth σ. While this weighting increases the importance of $f(x_{S_i}, x^*_{l,\bar{S}_i})$ proportional to the size of S, it however does not ensure that higher weights are assigned to model evaluations for observations closer to x.

A simple solution to the locality problem is to fit a local linear approximation in the form of a tangent line that predicts the black-box in a small neighbourhood around x, as in [67, 68, 72, 73]. Such an approach has however several drawbacks compared to SHAP (and thus Neighbourhood SHAP) such as higher instability, less interpretability, and assuming a fixed parametric form. While SHAP (and Neighbourhood SHAP) does not make any assumptions on the form of f in the feature space, local linear approximations assume linearity of the black-box in a neighbourhood. As a consequence, this may result in misleading attributions, as was demonstrated in Figure 3.1. See Supplement A.1 for a detailed discussion of local approximating models versus local reference populations.

3.3 Neighbourhood SHAP

Shapley values – similarly to other feature removal methods – employ a global reference distribution when computing attributions. This can lead to surprising artefacts as illustrated in Figure 3.2. To increase the local fidelity of Shapley values, we propose to sample from a well-defined local reference population instead. Having selected a distance metric D, such as the Euclidean distance or using a non-parametric regression like in *Bloniarz et. al* [74], we define a distance-based distribution $d : \mathbb{R}^m \to \mathbb{R}$ that is centred around x, such as the exponential kernel $d(x^*_{\bar{S}} \mid x_{\bar{S}}) = \exp(-D(x_{\bar{S}}, x^*_{\bar{S}})^2/\sigma^2_{nbrd})$. Further, we define the local neighbourhood distribution as

$$n(x^*_{\bar{S}} \mid x) = n_c \cdot d(x^*_{\bar{S}} \mid x_{\bar{S}}) \cdot r(x^*_{\bar{S}} \mid x)$$

where $r(x^*_{\bar{S}} \mid x)$ can be any marginal or conditional reference distribution and n_c is the normalising constant. This choice ensures that we sample neighbourhood values not only considering the metric space w.r.t. x but also the data distribution. This leads to a proposed change to the optimisation problem of eq. (3.1) to the following Neighbourhood SHAP

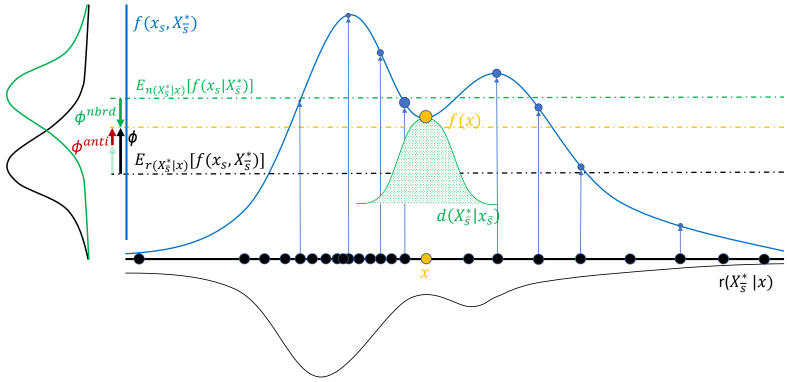


Figure 3.2: When sampling $\{x_i^*\}_{i=1}^L$ (black dots) from reference distribution $r(X_{\bar{S}}^* \mid x_S)$ (here $S = \emptyset$), the Shapley value ϕ at x is positive since $f(x)$ is larger than $\mathbb{E}_{r(X_{\bar{S}}^* \mid x)}[f(x_S, X_{\bar{S}}^*)]$. In contrast, Neighbourhood SHAP ϕ^{nbrd} is negative since $\mathbb{E}_{n(X_{\bar{S}}^* \mid x)}[f(x_S, X_{\bar{S}}^*)]$ is larger than $f(x)$. This difference results from the fact that, first, the model outcome has a local minimum at x, and second, $f(x_S, X_{\bar{S}}^*)$ takes its smallest values at the tails of the data distribution (right-skewed density of $f(x_S, X_{\bar{S}}^*)$ when $X_{\bar{S}}^* \sim p(X_{\bar{S}}^*)$, black line on the left). SHAP only captures that $f(x)$ is higher than the average model outcome but not that $f(\cdot)$ is smaller at x than it is for any other close observation – this is reflected by Neighbourhood SHAP.

minimisation

$$\mathop{\mathbb{E}}_{p(S)}\left[\left(\mathop{\mathbb{E}}_{n_c d(X_{\bar{S}}^* \mid x_{\bar{S}}) r(X_{\bar{S}}^* \mid x_S)}[f(x_S, X_{\bar{S}}^*)] - g(S)\right)^2\right].$$

Instead of estimating the neighbourhood distribution, which would require training an additional black-box, we approximate the expectation of the model outcome in the neighbourhood around x using self-normalised importance sampling [75] with proposal distribution $r(X_{\bar{S}}^* \mid x_S)$

$$\mathop{\mathbb{E}}_{n(X_{\bar{S}}^* \mid x)}[f(x_S, X_{\bar{S}}^*)] = \mathop{\mathbb{E}}_{r(X_{\bar{S}}^* \mid x_S)}\left[n_c \cdot d(X_{\bar{S}}^* \mid x_{\bar{S}}) f(x_S, X_{\bar{S}}^*)\right] \approx \frac{\sum_{i=1}^L d(x_{i,\bar{S}}^* \mid x_{\bar{S}}) f(x_S, x_{i,\bar{S}}^*)}{\sum_{i=1}^L d(x_{i,\bar{S}}^* \mid x_{\bar{S}})}.$$

While our proposal, Neighbourhood SHAP, weights the $f(x_S, x_{i,\bar{S}}^*)$ based on a distance metric to x, KernelSHAP uses uniform weights, i.e. $d(x_{\bar{S}}^* \mid x_{\bar{S}}) = 1$. We note that the proposed local neighbourhood sampling scheme has a convenient form which corresponds to the well-known Nadaraya-Watson estimator [76, 77, 78] used for kernel regression. Kernel regression is a non-parametric technique to model the non-linear relationship between a dependent variable Y (here, $f(x_S, X_{\bar{S}}^*)$) and an independent variable Z (here, $X_{\bar{S}}^*$), by approximating the conditional expectation $\mathbb{E}[Y \mid Z]$ (here, $\mathbb{E}_{r(X_{\bar{S}}^* \mid x_S)}[f(x_S, X_{\bar{S}}^*) \mid X_{\bar{S}}^*]$). While the form of the Nadaraya-Watson estimator has so far been justified from a kernel theory perspective (Supplement A.6), we show that it can be interpreted as an importance sampling estimator.

Proposition 1. *The Nadaraya-Watson estimator*

$$\widehat{\mathbb{E}}[Y|Z = z^*] = \frac{\sum_{i=1}^{L} d(z_i|z^*)y_i}{\sum_{j=1}^{L} d(z_j|z^*)},$$

where $d(z|z^*)$ is a kernel function, is a consistent self-normalised importance sampling estimator of $Y(Z)$ with proposal distribution $p(z)$ and desired distribution proportional to $p(z)d(z|z^*)$.

As pointed out in Supplement A.2, all Shapley axioms [57, 62] still hold true for the Neighbourhood SHAP. Now, by linearity, we can quantify the difference between SHAP and Neighbourhood SHAP as 'Anti-Neighbourhood SHAP' (see Supplement A.4). Looking at this difference might be of value to characterise the information loss when contrasting an instance to the global population instead of to a local neighbourhood. Finally, we also derive a variance estimator of Shapley values computed with the Shapley formula in Supplement A.10.

On-Manifold Explainability. A major disadvantage of marginal Shapley values and LIME is that the concatenated data vectors $(x_S, x^*_{i,\bar{S}})$ for a sampled reference x^*_i do not necessarily lie on the data manifold [61, 79]. This has two serious ramifications:

1. The model is evaluated in regions that lie off the data manifold where it might behave unexpectedly, and be unrepresentative for the data population; a similar problem was described by [80] who note that removal based methods induce bias in model explanations if removal is modelled by imputing with observations that are far from the actual test instance.

2. Adversaries can use an out-of-distribution (OOD) classifier trained to distinguish real-data from simulated concatenated data and, through this, construct a model whose Shapley values look fair even though the model is demonstrably unfair on the real-data domain [81].

To circumvent this problem, [69, 82, 83] propose the use of conditional instead of marginal reference distributions. However, using conditional reference distributions changes the interpretability of Shapley values – i.e. unrelated features get a non-zero attribution – and thus, their use is controversial [70]. A marginal Neighbourhood SHAP approach in contrast can achieve on-manifold explainability while keeping the properties of marginal Shapley values for small enough σ_{nbrd} if the data manifold is to some extent coherent (see Figure 3.3).

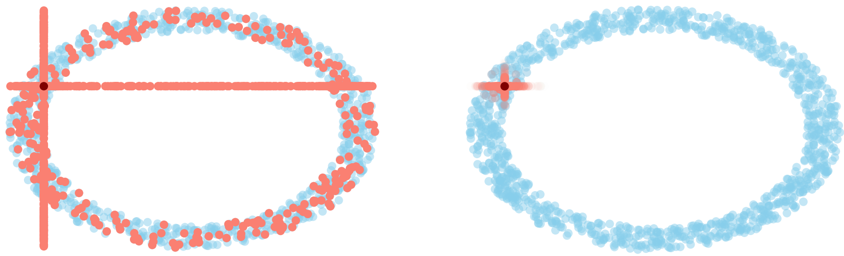


Figure 3.3: Concatenated data (pink dots) used for model evaluations for the computation of KernelSHAP (left) and Neighbourhood SHAP ($\sigma_{nbrd} = 0.1$, right) at a randomly sampled instance (maroon dots) where the data manifold is a ring in $\mathbb{R}^2$. Even though the background references (blue dots) lie on the data manifold, marginal Shapley values are evaluated at instances that lie off the data manifold.

Choice of Bandwidth. For $\sigma_{nbrd} \to \infty$, Neighbourhood SHAP will be equal to KernelSHAP, while it converges to 0 for $\sigma_{nbrd} \to 0$. Small neighbourhoods thus induce regularisation in the predictions which we also observe empirically in Section 3.5.

While SHAP values add up to $\sum_{j=1}^{m} \phi(j,x) = f(x) - \mathbb{E}_{r(X^* \mid x_S)}[f(X^*)]$, Neighbourhood SHAP attributions add up to $f(x) - \mathbb{E}_{n(X^* \mid x)}[f(X^*)]$. Hence, care needs to be taken when comparing SHAP and Neighbourhood SHAP, since the scales might differ.

In this case, both SHAP values (standard and neighbourhood) can be divided by either the sum of their absolute values or by their standard deviation, to represent relative attribution measures. As commonly observed with kernel regression approaches, there are some drawbacks, such as the additional hyperparameters (distance function, bandwidth) and increased variability especially in data sparse regions for small bandwidths. These problems can be tackled by choosing adaptive bandwidth methods. For instance, σ_{nbrd} could be chosen such that the 25% closest observations to x are not assigned more than 75% of the weight mass – thus making the explanation more local. We propose to plot the Neighbourhood SHAP values of the normalised features over a range of bandwidths, from $\sigma_{nbrd} = [0, 3m]$. This provides a powerful diagnostic and information tool.

The computational burden of changing σ_{nbrd} is not as large as it might first appear. Our importance sampling approach has the desirable property that $\mathbb{E}_{n(X^*_{\bar{S}} \mid x)}[f(x_S, X^*_{\bar{S}})]$ is estimated on the same set of references $\{x^*_l\}_{l=1}^{L}$ for each σ_{nbrd}, and that only the importance weights vary with the bandwidth. As a result, there are no additional model evaluations required when Neighbourhood SHAP is computed for a different σ_{nbrd}. This stands in contrast to other neighbourhood schemes proposed in the XAI literature such as KDEs [84], GANs [85] or Gaussian perturbations [86] where the black-box must be evaluated an additional $C \cdot L$ times for each new bandwidth where C denotes the number of sampled coalitions. Please refer to Supplement A.3 for a theoretical and empirical complexity analysis.

3.4 Smoothed SHAP

In the previous section, we discussed neighbourhood sampling as a useful tool to understand feature relevance through feature removal. We have also seen that the proposed neighbourhood sampling approach relates to kernel smoothers such as the Nadaraya-Watson estimator. This result can give us insights to consider a Smoothed SHAP that locally averages neighbouring SHAP values

$$\widehat{\phi}^{smtd}(j,x) = \frac{1}{\sum_{i=1}^{N} d^{smtd}(x'_i,x)} \sum_{i=1}^{N} d^{smtd}(x'_i,x)\widehat{\phi}(j,x'_i) \tag{3.2}$$

where $\{x'_i\}_{i=1}^{N}$ are samples from the reference distribution and d^{smtd} is a kernel function. Such smoothing procedures have been applied before in the explainability literature, e.g. for gradient-based methods [87, 88], and can be of interest when the interpretability of SHAP values suffers under the high instability of the black-box [89, 90, 91]. The smoothing it induces can be captured by a Lipschitz constant whose upper bound decreases with the bandwidth σ_{smtd}.

Theorem 2. *For every $x_0 \in \mathbb{R}^m$ with $||x - x_0|| < \delta$, there exists a constant $0 < L \leq max_y(f(y) - \mathbb{E}_{r(X^* \mid x_S)}[f(X^*)])h(\sigma_S^2)$ such that $||\widehat{\phi}^{smtd}(j,x) - \widehat{\phi}^{smtd}(j,x_0)|| \leq L||x - x_0||$ for the smoothed Shapley value estimator $\widehat{\phi}^{smtd}(j,x)$ (3.2) with $d(x,x_i) = \exp(-||x - x_i||^2/\sigma_S^2)$ if $f(\cdot)$ is bounded on $\{x_i\}_{i=1}^{N}$ where $h(\sigma_{smtd}^2)$ is a function that decreases in σ_{smtd}^2 with $h(\sigma_{smtd}^2) \to \infty$ as $\sigma \to 0$.*

Lipschitz continuity has been established as a favourable characteristic of explainability tools [89]. Introducing smoothness can indeed lead to feature attributions that do not correctly reflect the model's behaviour. In certain circumstances, this disadvantage can be outweighed by the advantage of more "intuitive" explanations [30] as these live closer to the global feature importance values.

With the tools from before, we can derive that Smoothed SHAP is an consistent importance sampling estimator of the SHAP values from the neighbourhood around x

$$\phi^{smtd}(j,x) = \mathop{\mathbb{E}}_{n(X' \mid x)}[\phi(j,X')] = \mathop{\mathbb{E}}_{p(S)}[v^{smtd}(S \cup \{j\},x) - v^{smtd}(S,x)] \tag{3.3}$$

where the new value function is defined by :

$$v^{smtd}(S,x) = \mathop{\mathbb{E}}_{n(X' \mid x)}\Big[\mathop{\mathbb{E}}_{r(X_{\overline{S}} \mid X'_S)}[f(X'_S, X^*_{\overline{S}})]\Big]$$

This smoothed value function relates to the explicit modelling of feature inclusion and gives an interesting perspective on the meaning of smoothing, namely that x is a measurement of the test instance variable X'. Exploring a smoothed summary of the SHAP values in the local neighbourhood around x highlights how local variability in $f(x)$ drives changes in the SHAP feature attributions. This is interesting in its own right but particularly so if features are susceptible to reporting error.

As an illustration, consider a black-box algorithm that predicts the fitness level of an adult based on multiple covariates, including weight. The reported weight may be subject to error if unreliable scales are used. In addition, as weight varies constantly throughout the day, the individual might not be interested in the attribution for one particular weight at a single point in time, but rather in the attribution that a range of weights per day receives.

The test instance is thus more appropriately described by a test *distribution* of X' around x where X' is a random variable that describes the volatility in the covariates of the test instance. If the test distribution is unknown, it can be estimated by setting it, as earlier, equal to a neighbourhood distribution $n^{smtd}(X' \mid x) \propto r^{smtd}(X' \mid x) d^{smtd}(X' \mid x)$ where $d^{smtd}(X' \mid x)$ encapsulates the prior belief on the variability of X' and $r^{smtd}(X' \mid x)$ captures the artefacts of the data distribution (i.e. skew, curtosis, high density regions). The kernel $d^{smtd}(x'_i, x) = \exp(-D^T(x'_i, x)\Sigma^{-1}_{smtd} D(x'_i, x))$ can now be defined with a multivariate bandwidth $\Sigma_{smtd} = \mathrm{diag}(\sigma^2_{smtd,1}, \ldots, \sigma^2_{smtd,m})$. We can observe empirically that such a choice can decrease the MSE of the estimation of Shapley values (Supplement A.11). Building upon results from kernel regression, we can quantify the squared distance of Smoothed SHAP to $\phi(j,x)$ (Supplement A.8). Finally, we also derive a variance estimator for Smoothed SHAP in Supplement A.10.

Choice of Smoothing Bandwidth. Prior information on the variability of the covariates of the test instance can be included in the definition of the bandwidth matrix. Fixed covariates, like age or season, are not expected to change and thus receive a bandwidth $\sigma_{smtd,j} \to 0$, while volatile features like weight, temperature or windspeed are assigned a positive bandwidth. For bandwidths $\sigma_{smtd,j} \to \infty$, the feature is treated as inherently missing. If $\sigma_{smtd,j} \to \infty$ for all features j, Smoothed SHAP equals the average of the Shapley values over all references which is often used as a global explanation measure [79, 83, 92]. As Smoothed SHAP can be estimated efficiently once SHAP values have been computed for the reference population, we propose, again, computing it for several bandwidth choices, and using a plot with respect to the bandwidth as a visualisation technique to help inform the choice of bandwidth. The bandwidth induces a bias-variance trade-off as derived in Supplement A.8: the larger the bandwidth, the smoother the results, but also the less Smoothed SHAP reflects the model behaviour at x, especially if f is highly non-linear.

Connection to LIME. Tabular LIME [22, 93] provides the same explanation for any two instances falling into the same quantile along each dimension [93]. As such it is also an aggregated attribution measure, similar to Smoothed SHAP. Key differences are the treatment of different dimensions and no proven guarantees of Lipschitz continuity (see Supplement A.5).

3.5 Examples

We present comprehensive experiments on several standardised real-world tabular UCI data sets [94] of different sizes predicted with ensemble classifiers or regressors, as well as an image classification task on the MNIST dataset. The experiments demonstrate some key attributes of Neighbourhood and Smoothed SHAP including: Neighbourhood SHAP increases on-manifold explainability and robustness against adversarial attacks; Neighbourhood SHAP also leads to sparser attributions than standard Shapley values; Smoothed SHAP tells us how Shapley values of neighbouring observations differ from the attribution of the test instance.

Since Neighbourhood SHAP, Smoothed SHAP and SHAP operate on different scales, we divide all attributions by their standard deviation (over features) unless otherwise specified. We present a subset of our results in this Section and refer the interested reader to Supplement A.11 for a thorough report of all experimental results (including simulated experiments), details and hyper-parameter settings.

On-Manifold Explainability and Robustness against Adversarial Attacks. For adversarial learning, we train a Random Forest and a LightGBM as OOD classifiers that distinguish true data from concatenated vectors used for model evaluations. We find that for small bandwidths σ_{nbrd}, the adversary is not able to distinguish between the test data and the concatenated test data (Figure 3.4), leading to a deterioration in their ability to discriminate true from concatenated vectors. Under the assumption that the classifiers are able to detect the true data manifold, we can thus claim that Neighbourhood SHAP

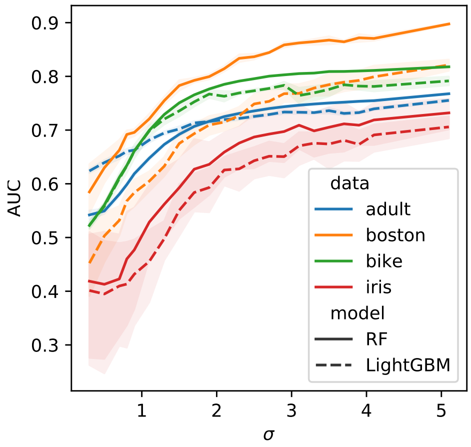


Figure 3.4: AUC from OOD LightGBM and RF over 10 runs with 95% CIs. Concatenated data was created by sampling as many coalition vectors as data and masking with random references. Where references are sampled locally (smaller σ_{nbrd}), OOD classifiers perform significantly worse.

relies more on observations from the data manifold than SHAP and LIME. Further, we mimicked the experimental setup of [81] on the COMPAS data set [15]: an adversarial black-box predicts recidivism based only on race if the data is predicted from the OOD classifier to be from the data manifold, and returns an unrelated column if it is not. As presented in Figure 3.5 for 10 randomly sampled individuals, the unrelated column has no effect on Neighbourhood SHAP and race has a higher relative attribution than it does for KernelSHAP.

Increased Local Prediction Performance. As SHAP learns a binary feature model $g(S) = \phi_0 + \sum_{j=1}^{m} \phi_j \mathbb{I}(j \in S)$, we can sample feature coalitions and reference values to perturb test data and predict the model outcome at the perturbed data. To check local prediction performance, we weight the MSE at each test instance and at each reference value with an exponential kernel of the distance between test instance and reference value. Its bandwidth signifies the size of the neighbourhood. Figure 3.6 presents the MSE corresponding to an XGBoost model, applied to four different datasets. As expected, Neighbourhood SHAP with a smaller bandwidth predicts data within a small neighbourhood significantly better than Neighbourhood SHAP with a larger bandwidth. Here we noticed that the difference between the bandwidths is larger where there are fewer features in the data set (such as the `iris` dataset). We attribute the loss in performance to the difficulty of estimating meaningful distances in high dimensions.

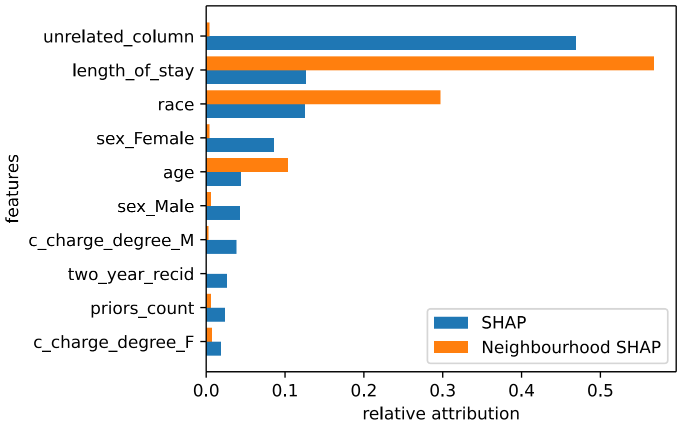


Figure 3.5: Adversarial black-box predicts recidivism using the COMPAS data. Absolute attributions obtained from Neighbourhood SHAP and KernelSHAP are divided by the sum of attributions for comparability. The adversarial attack affects Neighbourhood SHAP (with $\sigma_{nbrd} = 0.5$) less than KernelSHAP when averaged over 10 runs. Without adversarial attack, (Neighbourhood) SHAP attributes only `race` (not shown).

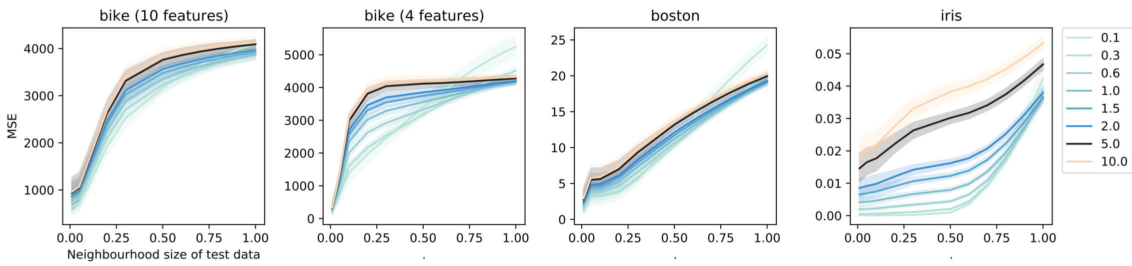


Figure 3.6: MSE when predicting local model outcome of an XGBoost model averaged over 400 runs displayed with 95% confidence intervals. Neighbourhood SHAP with smaller bandwidth predicts neighbourhoods significantly better than with large bandwidths.

Interpretation of Neighbourhood SHAP. Neighbourhood SHAP computed with small kernel widths reflects feature attributions when contrasting with model behaviour at similar observations, whereas Neighbourhood SHAP computed with large kernel widths renders model behaviour contrasting at a population scale. Figure 3.7 shows the evolution of Neighbourhood SHAP across bandwidths on randomly picked observations across different data sets. The test instance in the bike data set, where a XGBoost regressor predicts daily bike rentals, has a high normalised temperature of 0.82. As the median observation has a temperature of 0.50, the neighbourhood of our test instance is expected to look considerably different to the global population. For small kernel widths, Neighbourhood SHAP computes a negative attribution for temperature, whereas marginal SHAP is positive. This sign 'flip' is coherent with descriptive statistics: for a subpopulation with temperatures +/-0.05 around 0.82, temperature is negatively correlated with outcome (correlation equal to -0.08) whereas overall, bike rental tends to increase on warmer days (unconditional correlation equal to +0.47). Neighbourhood SHAP thus shows that a warmer temperature has in general a positive impact on the count of rental bikes, which reverses for very hot days. Standard Shapley values do not provide this type of fine-grained interpretation.

Similarly, in the Boston data set (see Figure 3.7), our test instance is a dwelling with a high percentage of lower status population (LSTAT) equal to 18.76%. LSTAT gets positive Neighbourhood SHAP values for small kernel widths, whereas its marginal Shapley value is negative. This observation is consistent with the negative overall correlation, which is equal to -0.76, whereas for a restricted population with LSTAT +/- 1% it is equal to +0.15. For similar dwellings i.e. with a high pupil-teacher ratio and many rooms, lower status populations do not decrease the value of the home as much as they do in general, and can even increase it.

As the median observation of the adult data set is middle aged with no capital gain, his neighbourhood is expected to look considerably different to the global population. The marginal Shapley value for age of nearly -0.2 tells us that setting the age of a randomly sampled person from the data set equal to 29 will lead to a drop of -0.2 in the predicted probability of high income (Figure 3.7). The Neighbourhood SHAP value drops however to -0.4 before it converges to -0.2. As a result, we know that setting the age of someone in the neighbourhood of the test instance (i.e. high capital gain observations) equal to 29 will lead to an even higher drop in the predicted probability of high income. In the boston data set, we see that the predicted price of a random house would decrease if we were to set its number of rooms to 6. However, houses that look similar to the test instance would be predicted considerably more valuable if their number of rooms was to bet set to 6. Again this is expected, since the test instance is a house in a low income region. Compared to houses in a similar neighbourhood, 6 rooms is quite large.

Interpretation of Smoothed SHAP. In contrast, Smoothed SHAP summarises marginal Shapley values (which contrast against the entire population) within a neighbourhood, instead of at a single instance. For example, consider the adult data set (Figure 3.7, first column). We chose a test instance for which the model performs poorly: its predicted probability of high income for this individual, aged 42, is equal to 0.09, when in actual fact the person has a high income. It is interesting to contrast the conventional Shapley value assigned to the person, which is obtained by Smoothed SHAP with a $\sigma_{smtd} \to 0$, with the average Shapley values for individuals like them.

We observe that Smoothed SHAP quickly assigns a negative attribution to age and a positive attribution to education for $\sigma_{smtd} > 1$, whilst SHAP values were positive and negative, respectively for the individual. This highlights local instability in the Shapley values, as the SHAP numbers for people similar to the predicted person are positive for education, and negative for the age feature. For the Boston data set we note that Smoothed SHAP of the Pupil/Teacher Ratio (PTRATIO) initially decreases for a small σ_{smtd}, as there are many dwellings with a high PTRATIO in the data neighbourhood of the test instance, while it then increases as the global attribution of this feature is in general higher.

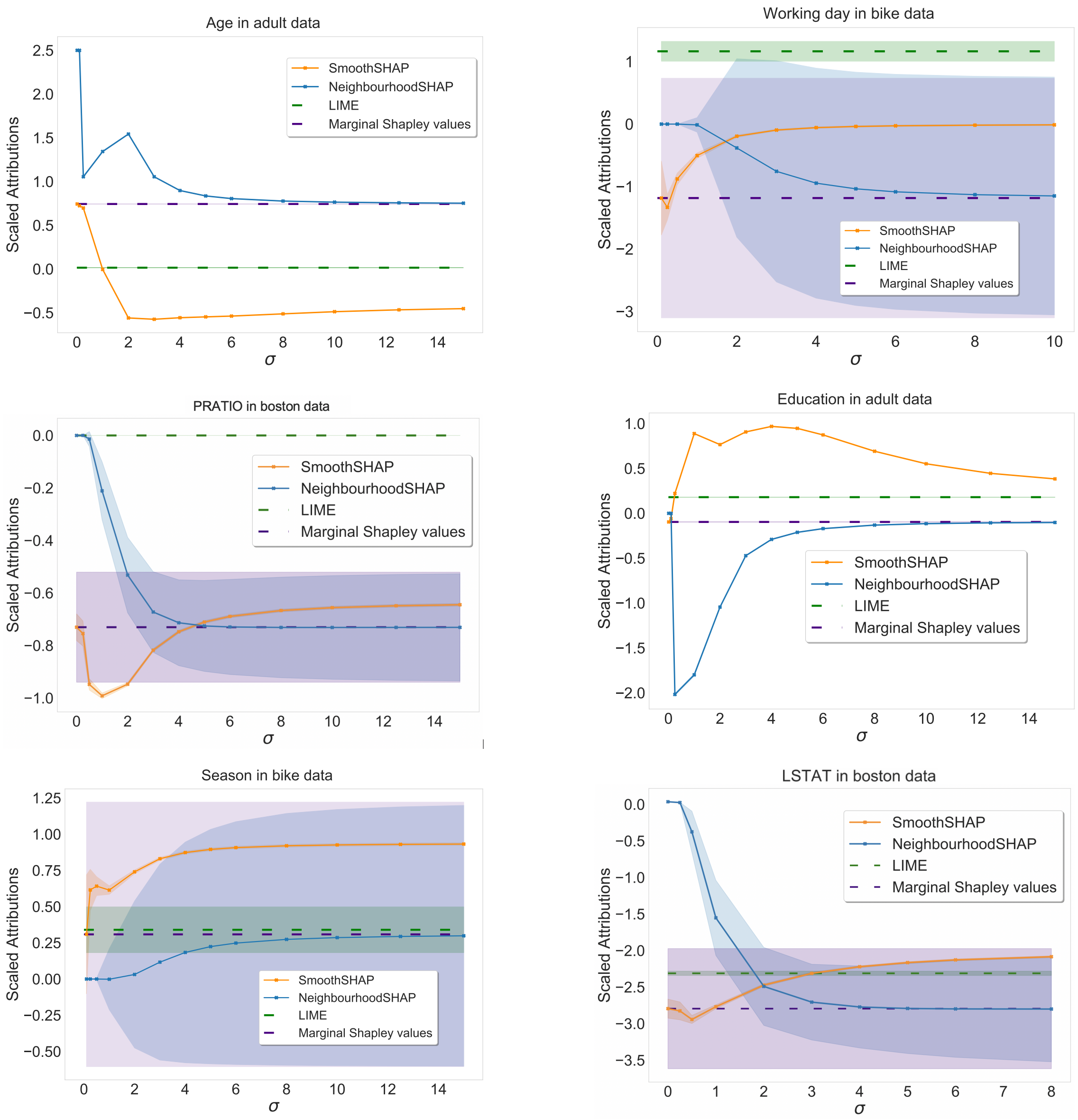


Figure 3.7: Scaled attributions at three different test instances (see Supplements A.11) for varying kernel widths computed with 2000 reference points in the adult, bike, and Boston housing datasets. Bounds for LIME have been computed over 2000 runs, while the Shapley bounds have been estimated with their theoretical formula as outlined in Supplements A.10.

Image Classification. We applied our Neighbourhood SHAP approach on KernelSHAP and also on DeepSHAP [57] which computes Shapley values for images based on gradients. After training a convolutional neural network on the MNIST data set, we explain digits with the predicted label 8 given a background data set of 100 images with labels 3 and 8.

As we see in Figure 3.8, Neighbourhood DeepSHAP gives pixels close to the strokes attributions with the highest absolute values while DeepSHAP assigns less sparse and more blurry attributions. This is expected: DeepSHAP compares each digit to a random digit in the population, while Neighbourhood DeepSHAP only looks at images in the neighbourhood, i.e. to images which have a similar stroke. As we show in the Supplement, the change in log odds of predicting a 3 after modifying the images depicting an 8 with the attributions (setting blue pixels to 0) is (non-significantly) higher for Neighbourhood DeepSHAP than it is for DeepSHAP.

In contrast, Smoothed DeepSHAP leads to a smaller change in log-odds which is expected since we lose information by smoothing. However we see that Smoothed DeepSHAP gives additional insights compared to DeepSHAP and Global DeepSHAP: in all images the lower left corner of the 8s is highlighted in blue only for Smoothed DeepSHAP. Thus, we know that there is at least one observation in the neighbourhood of these 8s that has a strong negative attribution in that image area. This image however loses importance when aggregating over the whole data set. Note that LIME gives the sharpest results because we chose the hyperparameters such that the image is split into the largest number of super pixels. We however see that LIME gives counter-intuitive results (i.e. lower right corner of the third 8 gets the lowest attribution, lower contour of first 8 gets highest attribution).

Another popular explanation method called Integrated Gradients (IG) [95] was added to Figure 3.8 for comparison purposes. This method consists in computing the gradients along a path from the input of interest to a baseline input. Thus it resembles feature-removal based approaches in so far as it requires baseline values to be specified. The baseline is typically defined to be just the mean of the feature. If a fixed baseline value is chosen, IG suffers under the same globality problem that was outlined for SHAP.

Evaluation of the proposed methods The two proposed methods were evaluated using a deletion metric from the ROAR framework developed by Hooker et. al [96]. For each observation, 20% of the most important features (measured by the magnitude of the absolute feature attributions) are imputed by their mean and the model is retrained. Eventually, we compare the resulting changes in predictive accuracy. We distinguish

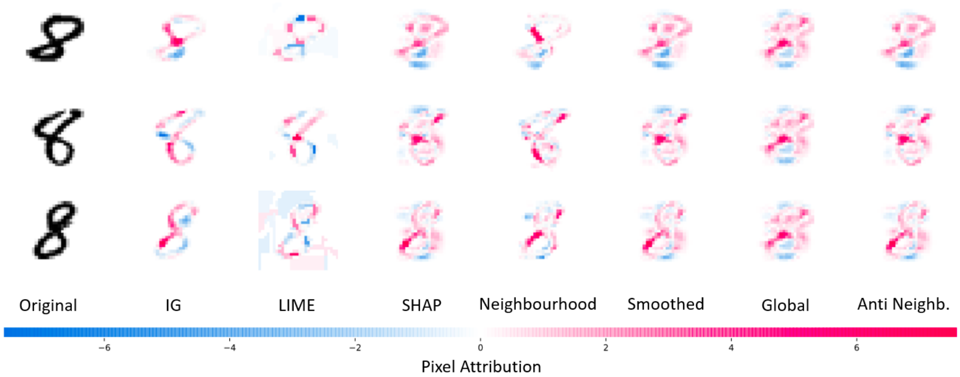


Figure 3.8: Randomly picked test images with explanations of the label 8. Red regions are pixels that increase the predicted probability of label 8 while blue regions decrease the predicted probability contrasted with the background data set.

between two settings: imputation with the (global) mean of the features and imputation with the local mean (i.e. the weighted mean of each feature where the weights were computed based on a distance metric to the observation). Results are shown in the Supplementary Material (see A.11.7). Overall, Neighbourhood and Smoothed SHAP result in a higher absolute change in test performance than Marginal SHAP.

3.6 Discussion

In this paper, we first highlighted the limitations of using SHAP when the local model behaviour is of interest. We then introduced Neighbourhood SHAP. While neighbourhood sampling has been applied in other areas of model explainability, such as image perturbations by adding noise [97], local linear approximations (see Supplements A.1), or rule-based models [98, 99, 100], it has not been previously introduced for model agnostic additive feature models such as SHAP.

Our contribution is important as it provides a theoretical understanding of explanations of local model behaviour, which is often lacking in the explainable AI literature [93]. A secondary contribution of this work is the analyses of how smoothing Shapley values can identify unstable feature attributions. While it is difficult to evaluate model explanations numerically, we provide an exhaustive comparison of different metrics (adversarial robustness, prediction accuracy, and visual inspection).

Neighbourhood SHAP and Smoothed SHAP both merit consideration, as they have considerable advantages compared to standard KernelSHAP. For comparability across experiments, we limited our analysis to the use of the euclidean distance as a distance metric.

In high dimensional spaces, this choice can be misleading [101] and the use of more powerful distance metrics, such as one obtained by random forests, would be appropriate. In future work, we may use adaptive kernels where the bandwidth isn't fixed, but shrinks in dense regions [102]. We thus caution against exclusively relying on mathematical metrics for explaining models, and suggest comparing the un-weighted and weighted histograms before any judgement calls. While it can be difficult to choose an adequate bandwidth, we see that having control over kernel width allows the user to have a precise understanding of model predictions, both locally and at a larger scale. LIME or KernelSHAP in their default implementation do not allow for such a detailed analysis.

Plots of Neighbourhood SHAP and Smoothed SHAP w.r.t. the bandwidths σ_{nbrd} and σ_{smtd} respectively are thus powerful tools that give additional insight into oblique dynamics of the black-box.

Limitations and Societal Impact Due to the increasing use of ML-based systems to assist or to replace the human decision-maker, there is a need for transparency in algorithms. Developing robust methods that provide end-users with insightful explanations on the decisions made is key for building assurance in critical systems employing ML. However, we acknowledge that as all explainability tools, our method requires good model fit. Finally, we caution against over-reliance on mathematical metrics for explaining models, as is no consensus evaluation tool for explainability models.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	On Locality of Local Explanation Models
Publication Status	X Published ☐ Accepted for Publication ☐Submitted for Publication ☐Unpublished and unsubmitted work written in a manuscript style
Publication Details	Ghalebikesabi, S.*, Ter-Minassian, L.*, DiazOrdaz, K. and Holmes, C.C., 2021. On locality of local explanation models. *Advances in neural information processing systems*, *34*, pp.18395-18407.

Student Confirmation

Student Name:	Lucile Ter-Minassian		
Contribution to the Paper	I am joint first author of this paper with Sahra Ghalebikesabi. I came up with the motivation to improve the locality of Shapley values. We collaboratively came up with the idea to use distance-based weighted sampling. Sahra derived that the Nadaraya-Watson estimator can be expressed as a self-normalised importance sampling estimator on her own. I implemented and ran most experiments but the one on adversarial attack (Figure 3.5). Jointly with Sahra, we wrote the paper. Prof. Holmes and Dr Diaz-Ordaz advised along the way and corrected the paper draft.		
Signature	LTM	Date	18/02/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments			
Signature	Chris Holmes	Date	25/02/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

PWSHAP: a path-wise explanation model for targeted variables

Abstract

Predictive black-box models can exhibit high accuracy but their opaque nature hinders their uptake in safety-critical deployment environments. Explanation methods (XAI) can provide confidence for decision-making through increased transparency. However, existing XAI methods are not tailored towards models in sensitive domains where one predictor is of special interest, such as a treatment effect in a clinical model, or ethnicity in policy models. We introduce Path-Wise Shapley effects (PWSHAP), a framework for assessing the targeted effect of a binary (e.g. treatment) variable from a complex outcome model. Our approach augments the predictive model with a user-defined directed acyclic graph (DAG). The method then uses the graph alongside on-manifold Shapley values to identify effects along causal pathways whilst maintaining robustness to adversarial attacks. We establish error bounds for the identified path-wise Shapley effects and for Shapley values. We show PWSHAP can perform local bias and mediation analyses with faithfulness to the model. Further, if the targeted variable is randomised we can quantify local effect modification. We demonstrate the resolution, interpretability, and true locality of our approach on examples and a real-world experiment.

4.1 Introduction

Recent years have seen an increase in the demand for transparency on machine learning-based decisions. In safety-sensitive settings particularly, practitioners need to understand how a model reasons in order to ensure its safe deployment in the future. In many scenarios, their attention focuses on assessing the importance of a specific predictor variable such as a treatment in a clinical model, or ethnicity with regards to model fairness. Ultimately, as humans naturally have a *causal* approach to model explainability, users may want to understand how the treatment[1] causally impacts the outcome i.e. through what mechanisms and if this matches their prior assumptions on the causal relationships in the data.

Here, we focus on the following question: in the presence of a general black-box ML model, how can we compute feature attributions according to the causal beliefs encapsulated by the posited DAG? We provide a framework for locally explaining a treatment's effect in such settings, with the following goals: *reliability*, *safety*, *interpretability* and *high resolution*.

Reliability, Safety, Interpretability in XAI An XAI model should aim at generating explanations that are *reliable*, *safe* and *interpretable*. Reliability, also known as being "true to the model", implies that the explanations do reveal the functional dependence of the model and are robust to distributional shifts. Safety relates to the ability to protect the framework from hazard, in particular attempts to fool models with deceptive data, also known as adversarial attacks. As defined in [24], interpretability is the degree to which a human can understand the cause of a decision.

High Resolution for Safety-Critical XAI In addition to the XAI goals cited above, resolution is often necessary when explaining a model to ensure its fairness. Following [103], we illustrate our point using a simple causal structure inspired by the Berkeley admissions dataset. Consider a predictive model for college entry with three features: sex, exam results and department. The sensitive attribute, sex, potentially impacts the predicted admission through both fair and unfair causal pathways. Sex may indirectly and fairly impact admission due to some individuals applying to more competitive departments. However, there may also be an effect through an unfair direct path, representing prejudice on the part of the admissions officer. This motivates *path-specific* measures of feature importance, instead of an overall single score that groups all paths together.

[1] We use the example of "treatment" in clinical models as illustrative of a central binary predictor variable.

Our Solution: Path-Wise Shapley (PWSHAP)
We introduce a method for explaining the local effect of a binary treatment under an assumed causal graph. We assume that (i) the treatment is an ancestor of the outcome in the directed acyclic graph (DAG) and that (ii) the DAG is compatible with the data, i.e. that it respects all conditional independences that could be found by running conditional independence tests in the data. The posited DAG may come from prior domain knowledge, or indeed be learnt from data and represents the user's beliefs. Our aim is not to understand the model's "internal DAG", and thus we do not assume that the posited DAG corresponds to the DAG of the underlying model. Instead, we aim to understand the black-box treatment effect as estimated by the model.

We show how augmenting the predictive model with such a causal DAG supports a novel targeted extension of SHAP values, allowing for the decomposition of the black-box treatment effect into interpretable path-wise Shapley effects. We provide stand-alone theoretical results for decomposing the original on-manifold Shapley value (i.e. Shapley with a conditional reference distribution) of a treatment feature into path-wise local causal effects. We claim that our method achieves the four goals presented above: reliability, safety, interpretability and high resolution.

Our contributions are as follows:

- We introduce Path-Wise Shapley (PWSHAP) effects, an extension of on-manifold Shapley values for locally explaining treatment effect under a causal DAG. Robustness to adversarial attacks (and thus safety) is guaranteed by the adoption of a conditional reference distribution. Reliability is ensured by the acknowledgment of the causal structure. As such, PWSHAP reconciles both safety and reliability.
- We show how our method can be used as a non-parametric alternative to mediation and bias analysis. We further show how PWSHAP can be used for fairness studies when the causal graph involves a mixture of fair/unfair paths, and under randomised treatment to assess effect modification (also referred to as moderation). We further show that Causal Shapley [104], the closest method to ours, does not acknowledge moderation.
- We establish error bounds (i) from the outcome model to the Shapley values and PWSHAP effects (ii) from the Shapley values and treatment model (referred to as propensity score) to the PWSHAP effects.

To the best of our knowledge, we are the first to interrogate the link between the Shapley feature importance of a treatment, and the standard notion of the treatment effect as defined within the Causal Inference literature, as the expected difference between potential outcomes under the two treatments.

4.2 Shapley values

Shapley values are a local feature attribution method. They quantify the importances of the features $\{1,\ldots,m\}$ of a complex machine learning model $f : \mathbb{R}^m \to \mathbb{R}^l$ at an instance $x \in \mathbb{R}^m$, given only black-box access to the model. The local prediction $f(x)$ is formulated as a sum of individual feature contributions: $f(x) = \phi_0^f(x) + \sum_{i=1}^M \phi_i^f(x)$, where $\phi_j^f(x)$ is the contribution of feature j to $f(x)$ and $\phi_0^f(x) = \mathbb{E}[f(X))]$ is the averaged prediction with the expectation over the observed data distribution.

The Shapley value of a feature j captures the change in model outcome comparing the prediction when the feature value x_j is included to when it's removed from the input. This change is computed from the difference in value function v when setting feature j equal to the instance feature value x_j, averaged over all possible coalitions S of features excluding feature j. If a feature is included in the coalition its value is set to the observed instance value x_j. To model feature removal, the value function takes the expectation of the black-box algorithm at observation x over the non-included features $\overline{S}$ using a reference distribution $r(X^* \mid x_S)$ such that

$$v_f(S,x) = \mathop{\mathbb{E}}_{r(X^* \mid x_S)}[f(x_S, X^*_{\overline{S}})]$$

for $\overline{S} := \{1,\ldots,m\} \backslash S$ and the operation $(x_S, x_{\overline{S}})$ denoting the concatenation of its two arguments.

Binomial weights $|S|!\,(m-|S|-1)!/(m-1)!$ take account of all possible orderings. The Shapley value of feature j is thus:

$$\phi_j^f(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [v_f(S \cup \{j\}, x) - v_f(S,x)],$$

$$\text{i.e. } \phi_j^f(x) = \mathbb{E}_{p(S \mid j \notin S)}[\phi_{j,S}^f(x)]$$

where $\phi_{j,S}^f(x) := v_f(S \cup \{j\}, x) - v_f(S,x)$ and $\forall j,\ p(S \mid j \notin S) = \frac{1}{m\binom{m-1}{|S|}}$.

Shapley values have become a gold standard amongst explanation models due to their desirable properties (model agnostic, additive) and axioms (*Symmetry*, *Efficiency*, *Linearity* and *Dummy*—see Supplement B.3.2 for details). However, the method has not been adopted in critical settings due to the considerable limitations of both possible reference distributions, marginal and conditional [61, 62, 70].

Limitations of Shapley Values On the one hand, *on-manifold* Shapley values [105] use a conditional reference distribution, conditioning on x_S to better account for correlations between features $r(X^* \mid x_S) := p(X^* \mid X^*_S = x_S)$. Sampling from a conditional distribution forces the model to be evaluated on plausible instances that lie on the data manifold. It thus improves the adversarial robustness and thus the safety of the method [81]. However, on-manifold Shapley values have been shown to be unreliable as they can generate misleading explanations [62, 70].

On the other hand, *off-manifold* Shapley values use a marginal reference distribution, that is $r(X^* \mid x_S) := p(X^*)$ [57]. The resulting explanations reveal the functional dependence better, also known as being "true to the model" [61]. However, sampling from the marginal distribution breaks the dependence between features. Consequently, off-manifold Shapley values are sensitive to adversarial robustness and thus deemed unsafe [81]. Note that adversarial robustness is key for fairness studies. If an unfair model undergoes an adversarial attack, it may "counterbalance" its potential prejudice on real-world data by forming predictions favourable to disadvantaged groups on implausible inputs. Since only the marginal distribution is used, the resulting Shapley value of a sensitive attribute might look fair even though the model would predict unfairly on real-world data [81]. Ultimately, Shapley values can't provide both reliable and safe explanations (i.e robust to adversarial attacks), which may hinder their adoption in safety-critical settings (see further details on Shapley values in Supplements B.3.1).

Shapley values also have limited interpretability. The attribution of a target feature j is the result of model evaluations averaged over all coalitions excluding j. The goal of this procedure is to acknowledge all the correlations amongst features. However, if some features are assumed to be independent, this assumption fails. Averaging over coalitions with/without independent features may generate redundancies and unbalance the resulting attribution. Also, the *interpretation* of on-manifold and off-manifold Shapley values is agnostic to the assumed *causal structure*, if any.

When a specific treatment is of interest, causal interpretation of its Shapley values should be done in light of the relative roles of other features: confounder, moderator or mediator; see Supplement B.1 for a definition of these notions. Interpreting Shapley values causally would be a case of the "Table 2 fallacy" [106], where all coefficients of a model are misleadingly interpreted as adjusted causal effects. Thus we claim that under a posited DAG, the Shapley value of a feature should be computed according to the assumed statistical dependencies, i.e. the *edges* in the DAG, and interpreted in light of its causal links with other variables, i.e. the *directions* of the arrows in the DAG.

In PWSHAP, we use a conditional reference distribution to ensure the robustness to adversarial attacks and safety of our method. Meanwhile, we are able to generate feature attributions that are both reliable and interpretable, thanks to the tailored causal interpretations of the effects we compute.

4.3 Path-Wise SHAP (PWSHAP)

The intuition behind the introduced method is two-fold. First, we decompose the Shapley value as a weighted sum of quantities that can be interpreted causally as treatment effects along coalitions. Second, by only considering relevant coalitions, we are able to deduce quantities that can be interpreted causally along paths. Since the treatment T is of special interest, we separate it from the other variables, that we call covariates C, such that $X = (C, T)$.

4.3.1 Problem Setup

Let C denote covariates, T a binary treatment and Y an outcome of interest. We assume that $Y = f^*(C, T) + \varepsilon$, with $\mathbb{E}[\varepsilon | C, T] = 0$. Our black-box f is an arbitrary function of $X = (C, T)$ which aims at predicting f^*.

We aim at explaining the specific effect of the treatment variable T on the predictions made by the black-box f for an individual with values c of covariates C. To do so, we first decompose the Shapley value of T into a weighted sum of "Shapley effects" which are inspired by conditional average treatment effects, commonly used in the causal literature. We refer to a coalition S excluding treatment T as a *subset of covariates* and note the value function as $v_f(S \cup \{T\}, c_S, t)$ when it is taken over the coalition $S \cup \{T\}$ and $v_f(S, c_S)$ when taken over S. Notations are summarised in Section B.2, with a running example to illustrate them all in Supplement B.8.

PWSHAP relies on two assumptions: (i) the treatment of interest is a causal ancestor of the outcome (no anti-causal learning) and (ii) the DAG is compatible with the observed data i.e. all conditional independence constraints implied by graphical d-separation relations hold in the data. The user-supplied DAG thus only encodes the conditional dependences and is not assumed to be identical to the underlying model behaviour. The "direction" of the arrows in the DAG is only used for causally *interpreting* the PWSHAP values.

4.3.2 Decomposition into Shapley Effects

First, we notice a connection between value functions v_f of the black-box f and conditional average treatment effects using coalition-wise Shapley values.

Definition 1 (Coalition-wise Shapley effect). We define the coalition-wise Shapley effect[2] of T on Y along the covariates C_S indexed by the subset of covariates S as:

$$\Psi^f_{T\to Y|C_S}(c_S) = v_f(S\cup\{T\}, c_S, 1) - v_f(S\cup\{T\}, c_S, 0)$$

The coalition-wise Shapley effect can be understood as a generalisation of conditional average treatment effects. Indeed, for the true model f^*, the RHS is equal to

$$\mathbb{E}[Y|C_S = c_S, T = 1] - \mathbb{E}[Y|C_S = c_S, T = 0]$$

Under the typical causal treatment effect identification assumptions, i.e. no interference, consistency, and conditional exchangeability given C [108], this is the conditional average treatment effect (CATE) [109] (definition in Supplement B.1) when S is the complete coalition, i.e. containing all covariates. In addition, $\Psi^f_{T\to Y|\emptyset}$ is the "base" treatment effect, i.e. a population-wide estimate of treatment effect. Its exact causal interpretation depends on the structure of the DAG, but in some cases it equates to the Average Treatment Effect (ATE) as defined by [109] (definition in Supplement B.1). The coalition-specific Shapley effect can be linked to the original Shapley values as follows.

Property 1 (Decomposing Shapley values into Shapley effects). The *coalition-wise Shapley value* $\phi^f_{T,S}(c,t)$ is equal to a weighted estimate of a local treatment effect,

$$\phi^f_{T,S}(c,t) = w^*_S(c_S,t)\cdot \Psi^f_{T\to Y|C_S}(c_S), \tag{4.1}$$

where $w^*_S(c,t)$ denotes what we call the "propensity weights" defined by

$$w^*_S(c,t) = t - p(T = 1|C_S = c_S)$$

. This name follows the fact that these weights are related to whether the sample is an outlier or not.

[2]Note that our Shapley effects are orthogonal to those introduced by [107] for numerical models .

The proof can be found in Supplement B.10.1. Property 1 shows that each coalition-specific term in the original on-manifold Shapley value is equal to the product of two terms. The first is a weight that depends on the propensity score. The second is a measure of the treatment effect, namely the coalition-specific Shapley effect. As a result, the overall Shapley value $\phi_T^f(c,t)$ can be decomposed as

$$\phi_T^f(c,t) = \mathbb{E}_{p(S|T\notin S)}[w_S^*(c_S,t)\cdot \Psi^f_{T\to Y|C_S}(c_S)].$$

4.3.3 Path-Wise Shapley (PWSHAP) Effects

Although we connected Shapley values to coalition-wise Shapley effects the latter still only apply to *coalitions* and not specific *paths*. However, the coalition-wise Shapley effect $\Psi^f_{T\to Y|C_S}(c_S)$ can be understood as the causal flow from T to Y through a set of covariates S. Thereby, we define the causal flow along the (undirected) path from T to Y through C_i as the difference between the causal flow through all covariates and the causal flow through all covariates but C_i. See Supplement B.6 for a generalisation of paths of length 3 or more.

Definition 2 (Path-wise Shapley effect). Let S^* be the coalition with all covariates. We refer to the following quantity as the path-wise Shapley effect of T on Y along the path from T to Y through C_i:

$$\Psi^f_{C_i}(c) = \Psi^f_{T\to Y|C_{S^*}}(c) - \Psi^f_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}}).$$

For instance in the admission example from Section 4.1, the path-wise Shapley effect of sex on admission (Adm) mediated by the chosen department (Dpt) $\Psi^f_{Sex\to Dpt\to Adm}$ is $\Psi^f_{Sex\to Adm|Dpt,Exam} - \Psi^f_{Sex\to Adm|Exam}$.

Path-wise Shapley effects thus quantify the change in model outcome when specifying the feature values along a specific path, compared to when all features are specified but the ones on the path of interest. As such, PWSHAP measures the effect of the treatment on the outcome through a causal pathway. Ultimately, conditioning on all other features reinforces the locality of our result. It can also be seen as a contribution to the shift from a global estimated "base" treatment effect to an individual estimated treatment effect. However, note that PWSHAP violates the efficiency property i.e. they do not sum up to an interpretable quantity like the original Shapley feature attributions do. Moreover, Property 2 shows that integrating PWSHAP effects can help isolate covariates that are conditionally independent on the treatment given other covariates (the Supplement B.10.2 for the proof). As shown in Section 4.6, the actual causal meaning of this conditional independence depends on the posited DAG of the data, however.

Property 2 (Integration of the PWSHAP effects). Let C_i be a covariate such that $C_i \perp\!\!\!\perp T|C_{-i}$ where $C_{-i} := C_{S^*\setminus\{i\}}$. Then for any function f and any value c_{-i} of C_{-i},

$$\mathbb{E}[\Psi^f_{C_i}(C_i, c_{-i})|C_{-i} = c_{-i}] = 0$$

4.3.4 Estimation of Shapley Effects from Shapley Values

Using Property 1, we can express the coalition-wise Shapley effects $\Psi^f_{T\to Y|C_S}$ from the coalition-wise Shapley values $\phi^f_{T,S}(c,t)$ as $\Psi^f_{T\to Y|C_S}(c_S) = \phi^f_{T,S}(c,t)/w^*_S(c,t)$. Therefore, the path-wise Shapley effects $\Psi^f_{C_i}$ are computed as:

$$\Psi^f_{C_i}(c) = \frac{\phi^f_{T,S^*}(c,t)}{w^*_{S^*}(c,t)} - \frac{\phi^f_{T,S^*\setminus\{i\}}(c,t)}{w^*_{S^*\setminus\{i\}}(c,t)}.$$

In practice, path-wise Shapley effects are computed by replacing the true propensity weights with weights that use an estimate of the propensity score. For this, we further need to assume positivity holds.

The path-wise Shapley effect of T on Y through C_i is thus estimated in three steps:

1. computing the coalition-wise Shapley values for S^* the entire set of covariates and $S^*\setminus\{i\}$;
2. dividing each of these terms by an estimate of their corresponding propensity weight;
3. taking the difference between the two resulting quantities (also known as coalition-specific Shapley effects) to isolate the effect along the path through C_i.

Note that division by weights requires overlap, that is $\forall c, 0 < p(T = 1|C = c) < 1$.

4.4 Related Work

4.4.1 Conceptual Distinction Between Local Explanations Models and Causal Inference

Causal Inference aims at assessing the effect of a *feature* at a global scale (e.g. ATE) or within subgroups (e.g. CATE). Contrastingly, Shapley values assess the *local* effect of a *feature value* compared to the values taken by that feature in the reference distribution. Therefore, to identify path-wise local effects, we consider a path to be "deactivated" when the covariate value gets sampled from the reference distribution, i.e. the covariate is *not* in the coalition.

Conversely, specifying a covariate value, i.e. when the covariate *is* in the coalition, "activates" a path. Thereby, coalition-wise effects are conditional treatment effects marginalised over covariates that aren't in the coalition. In the admission example from Section 4.1, the coalition-specific Shapley effect for $\{Exam\}$, $\Psi^f_{Sex \to Adm|\{Exam\}}$ corresponds to the treatment effect along two paths: the direct path $Sex \to Adm$ and the path from Sex to Adm through $Exam$.

4.4.2 Comparison of PWSHAP with Existing Methods

We compare our method to two baseline explanation methods.

Our first baseline is Causal Shapley (CS) [104], another method aiming to explain a model under an assumed causal DAG. Like PWSHAP, Causal Shapley splits Shapley attributions, although the split is binary (direct/indirect effect). In Causal Shapley, the indirect effect of a feature j, the distribution of the 'out-of-coalition' features changes due to the do-operator (see Suppl. B.3.1 and B.3.3 for further details).

Our second baseline is on-manifold Shapley, a natural choice given that PWSHAP augments the original method. Section B.3.4 details other graph based Shapley methods [110, 111], which are not appropriate baselines here due to structural differences.

Higher Model Fidelity, Lower Reliance on Causal Assumptions than Causal Shapley
We claim that PWSHAP has higher model fidelity and relies less on the assumed causal structure than Causal Shapley. As the direct/indirect effect split is based on *do*-calculus in Causal Shapley, the computation of the attributions depends on the assumed DAG (both the edges and their directions). In contrast, PWSHAP computations only depend on the hypothesised feature *dependencies* i.e. the edges in the DAG. Only the causal *interpretation* of PWSHAP depends on the direction of the edges. We view the fact that our approach is agnostic to the choice of a (compatible) DAG as a strength, as it allows different experts to explain the black-box model output according to their own causal beliefs about the data or phenomenon being studied (see B.3.6 for a detailed discussion on this). Ultimately, by applying *do*-calculus, Causal Shapley computes feature attributions according to preconceptions of how the model should reason, and as such is "forcing" explanations to fit to a presumed causal structure.

To further illustrate the limitations of relying on the causal assumptions and show that PWSHAP has higher fidelity to the model, let us consider a black-box with a single covariate

C, and a treatment T. If we wrongly assume C to be a confounder instead of a mediator, the indirect effect of treatment i.e. the mediation of treatment through C would be null according to Causal Shapley (see Property 6 in Supplement B.3.3). By contrast, only the causal interpretation of the PWSHAP effect through C would be incorrect, but its value would remain unaltered. Again, the *value* of an effect remains essential in fairness studies where we may want to check if a path plays any role in the decision-making process, regardless of the interpretation of this pathwise effect.

Increased Resolution, Better Interpretability Compared to both Causal and on-manifold Shapley, PWSHAP has higher resolution—as it is path-specific instead of feature-specific—and improved interpretability. In Causal Shapley and on-manifold Shapley values, feature attributions result from taking an average over coalitions, whereas PWSHAP only considers coalitions used to compute effects. Ultimately, evidence has shown that on-manifold Shapley values and Causal Shapley values can generate misleading interpretations [62]. In on-manifold Shapley values the attribution of a feature that does not appear in the algebraic formulation of the model can be non-zero, depending on how the data is distributed. This is induced by both the conditional reference distribution, the average taken over multiple coalitions. By providing an exact interpretation for the computed quantities, PWSHAP overcomes this unreliability issue.

If a PWSHAP effect $\Psi^f_{C_i}$ is null, it means that specifying the covariate $C_i = c_i$ has had no impact on the treatment effect compared to marginalising it, *according to our black-box* (see Lemma 1 and Property 5). Meanwhile, PWSHAP remains robust to adversarial attacks, as it samples from a conditional reference distribution [81]. PWSHAP thus reconciles *safety* and *reliability*. However, a limitation of PWSHAP compared to both baselines is that it violates the efficiency property (see Suppl. B.3).

4.5 Error bounds

We now show how to obtain error bounds for quantities like path-wise Shapley effects from other quantities like the outcome model, according to Figure 4.1. To the best of our knowledge, these are the first results regarding error bounds for on-manifold Shapley values. In the following, $(\hat{f}_N)$ denotes a sequence of estimators of f^*. The proofs can be found in Supplements B.10.3 and B.10.4

Property 3 (Convergence of the outcome model implies convergence of Shapley values and PWSHAP effects). If we assume $\forall c,t,N$ we have $|\hat{f}_N(c,t) - f^*(c,t)| \leq e_{\mathrm{N}}^{\mathrm{outcome}}$ then:

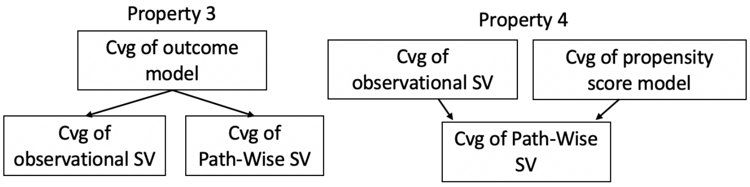


Figure 4.1: Error Bound infography (Cvg = Convergence, SV = Shapley value)

1. Convergence of all coalition-specific Shapley effects to the true coalition-specific Shapley effects:

$$\forall c, N, S \text{ s.t. } T \notin S, \qquad |\Psi^{\hat{f}_N}_{T \to Y|C_S}(c) - \Psi^{f^*}_{T \to Y|C_S}(c)| \leq 2e^{\text{outcome}}_{\text{N}}.$$

2. Convergence of the coalition-specific Shapley terms:

$$\forall c, t, N, \qquad |\phi^{\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)| \leq 2e^{\text{outcome}}_{\text{N}}.$$

3. Convergence of the path-wise Shapley effects:

$$\forall i, c, N, \qquad |\Psi^{\hat{f}_N}_{C_i}(c) - \Psi^{f^*}_{C_i}(c)| \leq 4e^{\text{outcome}}_{\text{N}}.$$

Property 4 (Convergence of estimated coalition-specific Shapley values and propensity score implies convergence of estimated path-wise Shapley effects). We assume that the arbitrary propensity score model π^N and the true propensity score model π^* verify ε-strong overlap, i.e. $\varepsilon \leq \pi^N \leq 1-\varepsilon$ and $\varepsilon \leq \pi^* \leq 1-\varepsilon$.
Further, if we assume:
$\forall c, N$ that $|\pi^N(c) - \pi^*(c)| \leq e^{\text{propensity}}_N$ and;
$\forall S$ s.t. $T \notin S, c, N$ that $|\hat{\phi}^{N,\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)| \leq e^{\text{Shapley}}_N$ then, defining

$$w^N_S(c,t) = t - \mathbb{E}_p(C_{\bar{S}}|C_S = c_S)[\pi^N(c_S, C_{\bar{S}})]$$

We have the following results:

4. Convergence of the estimated coalition-specific effects to the true coalition-specific effects:

$$\forall S \text{ s.t. } T \notin S, c, t, N, \quad |\hat{\Psi}^{N,\hat{f}_N}_{T \to Y|C_S}(c) - \Psi^{f^*}_{T \to Y|C_S}(c)| \leq \frac{2e^{\text{Shapley}}_{\text{N}}}{\varepsilon} + \frac{2||f^*||_\infty \cdot e^{\text{propensity}}_{\text{N}}}{\varepsilon^2}.$$

5. Convergence of the estimated PWSHAP effects to the true PWSHAP effects:

$$\forall i, c, t, N, \quad |\hat{\Psi}^{N,\hat{f}^N}_{C_i}(c) - \Psi^{f^*}_{C_i}(c)| \leq \frac{4e^{\text{Shapley}}_N}{\varepsilon} + \frac{4||f^*||_\infty \cdot e^{\text{propensity}}_N}{\varepsilon^2}.$$

4.6 Causal Interpretations of “Building Blocks” DAGs

PWSHAP effects are interpreted by revisiting Causal Inference concepts of confounding, moderation and mediation at a local scale. As our method stands on theoretical grounding, we first provide objective evidence using explicit equations.

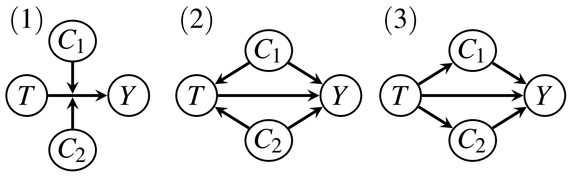


Figure 4.2: DAG for the building blocks

4.6.1 Local Bias Analysis

Under DAG (2) of Figure 4.2, PWSHAP effects are causally interpreted as follows:

$$\begin{aligned}\phi_T^{f^*} &= {}^{w_{12}^*}\!/_3 \cdot \mathrm{CATE}(c_1,c_2) + {}^{w_1^*}\!/_6 \cdot \text{ "CATE"}_{C_1}(c_1)\\ &+ {}^{w_2^*}\!/_6 \cdot \text{"CATE"}_{C_2}(c_2) + {}^{w^*}\!/_3 \cdot \text{Diff. in means}\end{aligned}$$

where $\text{"CATE"}_{C_S}(c_s) = \mathbb{E}[Y|T=1, C_S=c_S] - \mathbb{E}[Y|T=0, C_S=c_S]$. We refer to these terms as “CATE”s, in an abuse of notation, but note that they are not true causal conditional average treatment effects, as they include confounded paths. The term “Diff. in means” stands for $E[Y|T=1] - E[Y|T=0]$. To isolate the spurious effect of the confounders, we further assume that the two confounders are not effect moderators.

Definition 3 (Local confounding effect). In this example, we call the PWSHAP effect of C_2 the “local confounding effect of C_2” and note it $\Psi^f_{T\leftarrow C_2\rightarrow Y}$. In other words, $\Psi^f_{T\leftarrow C_2\rightarrow Y} := \Psi^f_{C_2}$

Notably, for the true model f^*, $\Psi^{f^*}_{T\leftarrow C_2\rightarrow Y}(c_1,c_2) = \mathrm{CATE}(c_1,c_2) - \text{"CATE"}_{C_1}(c_1)$. This quantity has been referred to as the bias due to unmeasured confounding (assuming we observe C_1 but not C_2) in a segment of the sensitivity analysis literature [112]. Therefore, our measure of confounding effect is a local equivalent of this bias. Indeed, integrating out this difference over C_1, C_2 yields $\mathrm{ATE} - \mathbb{E}[\mathbb{E}[Y|T=1,C_1] - \mathbb{E}[Y|T=0,C_1]]$ where ATE is the Average Treatment Effect. However, if a covariate is both a confounder and an effect modifier, its path-wise attribution will cover both phenomena and the two effects will be indiscernible. Ultimately, PWSHAP contrasts with sensitivity analysis methods which are meant for quantifying unobserved confounding, whereas our method measures the impact of an observed confounder. Further details on such techniques can be found in the Supplement B.3.7.

Lemma 1 (Integration of the local confounding effect, true model). Let C_1, C_2 be two pre-treatment covariates such that ignorability given C_1, C_2 holds, i.e. $\forall t, Y(t) \perp\!\!\!\perp T | C_1, C_2$. If, additionally, C_2 is not a confounder, i.e. C_1 alone guarantees ignorability or $\forall t, Y(t) \perp\!\!\!\perp T | C_1$, then the integral of the local confounding effect of f^* w.r.t. C_2 on the joint distribution of covariates is null:

$$\mathbb{E}[\Psi^{f^*}_{T \leftarrow C_2 \rightarrow Y}(C_1, C_2)] = 0.$$

The proof can be found in Supplement B.10.5. For a variable that is not actually a confounder, the integration of the local confounding effect thus yields zero. This can be generalised to any number of confounding pre-treatment covariates, by grouping all of them in C_1. For any blackbox f, a stricter condition yields the same result as a corollary of Proposition 2. We give that result and an example of local bias analysis in Supplement B.4.2.

Further, if the local confounding effect is zero for all individuals in the training set, then we can hypothesise that the model did not learn to predict through the confounding path $T \leftarrow C_2 \rightarrow Y$. Such a step is essential if we wish to understand how our model reasons. Consider a spurious path we do not want the model to predict from—i.e. we want to prevent "shortcut learning". We can do so by checking that all confounding effects are null, or using a diagnostic plot of all local confounding effects in the cohort.

4.6.2 Local Moderation Analysis Under Randomised Treatment

Here, "moderation" refers to an effect modification as in [113]. In the setting represented in Figure 4.2, causal graph (1), where treatment is assumed to be unconfounded, we interpret the PWSHAP decomposition as follows:

$$\begin{aligned}\phi^{f^*}_T &= 1/3 \cdot w^*_{12} \cdot \mathrm{CATE}(c_1, c_2) + 1/6 \cdot w^*_1 \cdot \mathrm{CATE}_{C_1}(c_1) \\ &\quad + 1/6 \cdot w^*_2 \cdot \mathrm{CATE}_{C_2}(c_2) + 1/3 \cdot w^* \cdot \mathrm{ATE}\end{aligned}$$

Definition 4 (Local moderating effect). In this example, we call the PWSHAP effect of C_2 the "local moderating effect of C_2" and denote it $\Psi^f_{C_2:T\rightarrow Y}$. In other words, $\Psi^f_{C_2:T\rightarrow Y} := \Psi^f_{C_2}$.

PWSHAP assesses the local effect modification induced by C_2 by "unspecifying" this feature. Having null local moderating effect would mean that C_2 did not act as a moderator for this specific subject, *according to our fitted black-box.* Unlike previous methods [40, 114, 115], our PWSHAP approach to moderation analysis does not involve subgroup finding—a technique known to be under-powered [116]—and is non-parametric (see Supplement B.3.7 for a review of moderation analysis). Ultimately, we show that in the presence of pre-treatment moderators, Causal Shapley compounds the main effect of treatment and its effect via moderation into a single "direct" effect, whereas PWSHAP explanations are able to distinguish the added treatment effect due to moderation from the main effect.

We compare PWSHAP with Causal Shapley on an example as shown in DAG (1) of Figure 4.2 assuming $Y = \beta T + \gamma_1 C_1 + \gamma_2 C_2 + \alpha_1 T C_1 + \alpha_2 T C_2 + \varepsilon$ with $\mathbb{E}[\varepsilon|T, C_1, C_2] = 0$ and where C_1, C_2 are two independent moderators with $C_1, C_2 \sim \text{Uniform}(0,1)$. Treatment is randomised: $T \sim \text{Bernoulli}(p)$. Details about the following are given in Supplement B.4.1. PWSHAP yields:

$$\begin{aligned}
&\Psi^{f^*}_{T\to Y|C_1,C_2} = \beta + \alpha_1 c_1 + \alpha_2 c_2 \\
&\Psi^{f^*}_{C_1} := \Psi^{f^*}_{C_1:T\to Y} = \alpha_1(c_1 - {}^1\!/_2) \\
&\Psi^{f^*}_{T\to Y|\emptyset} = \beta + {}^{\alpha_1}\!/_2 + {}^{\alpha_2}\!/_2 \\
&\Psi^{f^*}_{C_2} := \Psi^{f^*}_{C_2:T\to Y} = \alpha_2(c_2 - {}^1\!/_2)
\end{aligned}$$

where $\mathbb{E}[C_1] = \mathbb{E}[C_2] = {}^1\!/_2$. PWSHAP effects through C_1 and C_2 are null if $C_1 = C_2 = {}^1\!/_2$. The PWSHAP approach thus matches the default behaviour of local explanation methods: paths through effect moderators are given zero attribution if the moderator value is equal to the population average. This highlights the true locality of our method. Furthermore, one can check that the moderating effects integrate to 0. Again, this is coherent with the overall definition of randomised treatment in Causal Inference. By contrast, moderation by C_1 and C_2 is overlooked in Causal Shapley as $\phi^{f^*,\text{CS}}_{T,\text{indirect}} = 0$. Further, $\phi^{f^*,\text{CS}}_{T,\text{direct}} = (t-p)\{\beta + \frac{\alpha_1}{2}\cdot(c_1 + \frac{1}{2}) + \frac{\alpha_2}{2}\cdot(c_2+\frac{1}{2})\}$ which does not reflect the local behaviour of the model.

4.6.3 Local Mediation Analysis

Under DAG (3) of Figure 4.2, i.e. with unconfounded treatment and two mediators only depending on it, the causal interpretation of the PWSHAP approach to mediation is:

$$\begin{aligned}
\phi^{f^*}_{T} &= {}^1\!/_3 \cdot w^*_{12}\text{CDE}_{C_1,C_2}(c_1,c_2) + {}^1\!/_6 \cdot w^*_2\text{CDE}_{C_2}(c_2) \\
&\quad + {}^1\!/_6 \cdot w^*_1\text{CDE}_{C_1}(c_1) + {}^1\!/_3 \cdot w^*\text{ATE}
\end{aligned}$$

where CDE refers to the *Controlled Direct Effect* (definition in Supplement B.1), with $\text{CDE}_{C_S}(c_s) = \mathbb{E}[Y|T=1, C_S = c_s] - \mathbb{E}[Y|T=0, C_S = c_s]$. We claim that the difference in CDE is able to isolate the local effect of a given mediator and has a causal interpretation, as outlined by the local mediating effect introduced below.

Definition 5 (Local mediating effect). Here, we call the PWSHAP effect of C_2 the "local mediating effect of C_2" and note it $\Psi^f_{T\leftarrow C_2 \rightarrow Y}$. So, $\Psi^f_{T\rightarrow C_2 \rightarrow Y} := \Psi^f_{C_2}$.

Property 5 (Ancestors of outcome). Let M_1, M_2 be two post-treatment and pre-outcome variables. Assuming that variables C include all confounders of the relationships between T, Y and (M_1, M_2) and that $M_2 \perp\!\!\!\perp T, M_1 | C$, then for any value c of C and m_1 of M_1,

$$\mathbb{E}[\Psi_{T\rightarrow M_2 \rightarrow Y}(c, m_1, M_2) \mid C = c] = 0.$$

In other words, if M_2 is not mediating the effect of T on Y because $M_2 \perp\!\!\!\perp T|C$, and M_2 is independent of M_1 conditionally on T, C, then integrating the local mediating effect of M_2 yields 0 which is coherent with our intuition. The proof can be found in Suppl. B.10.6. See Suppl. B.3.7 for further comparisons with the traditional Natural Effects approach (definition in Supplement B.1) and with Causal Shapley.

4.7 Experiments: Synthetic Data

In the following two experiments, we show PWSHAP's ability to capture confounding and mediation on synthetic datasets. We infer path-specific Shapley effects following the procedure from Section 4.3.4 and compute the absolute values of their averages $\bar{\Psi}$ across the testing set. We divide these values by the empirical standard deviation of outcome σ_Y on the training set to mitigate variation due to the scale of the outcome. We follow the same process for Causal Shapley's direct and indirect effects w.r.t. the treatment. Results are averaged over 25 randomly sampled datasets, with standard errors shown in parentheses. More details are given in Supplement B.7.

Local Bias Analysis. We consider the previous model with two pre-treatment covariates C_1 and C_2 described in DAG (2) of Figure 4.2, and with results derived in Section 4.6.1. We look at three scenarios: (i) neither C_1 nor C_2 are confounders, (ii) C_1 is a confounder but C_2 is not, (iii) both are confounders.

Results are shown in Table 4.1. Local confounding effects are significantly higher for confounding variables compared to non-confounding variables. This shows how these effects can isolate individual confounders in pre-treatment covariates, in accordance with Lemma 1. Conversely, we do not notice any significant change in Causal Shapley's direct effect. Causal Shapley's indirect effect is even lower - as it is expected to be zero from Property 6.

Scenario	$\frac{\lvert\bar{\Psi}^f_{T\leftarrow C_1\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\Psi}^f_{T\leftarrow C_2\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{direct, CS}}_T\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{indirect, CS}}_T\rvert}{\sigma_Y}$
C_1, C_2 non-conf.	0.057 (0.011)	0.064 (0.013)	0.069 (0.010)	0.006 (0.001)
C_1 conf., C_2 not	0.505 (0.057)	0.054 (0.009)	0.067 (0.008)	0.002 (0.000)
C_1, C_2 conf.	0.322 (0.037)	0.277 (0.034)	0.076 (0.010)	0.006 (0.003)

Table 4.1: Results on local bias analysis.

Local Mediation Analysis. We consider DAG (3) of Figure 4.2, applied to a college admission example where we investigate the effect of sex—noted T — on the *logit* of the probability of admission, mediated by exam results $C_1 = Q$ and department choice $C_2 = D$. Details are in Supplement B.4.3. We look at three scenarios : (i) neither is a mediator, (ii) the former is a mediator but the latter is not, (iii) both are mediators.

Results are presented in Table 4.2. Local mediating effects are significantly higher for mediating variables compared to non-mediating variables. This shows that these effects can isolate individual mediators in post-treatment covariates, in accordance with Property 5. Conversely, Causal Shapley's indirect effect seems to capture the presence of mediators - but not which variables are mediators.

Scenario	$\frac{\lvert\bar{\Psi}^f_{T\rightarrow Q\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\Psi}^f_{T\rightarrow D\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{direct, CS}}_T\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{indirect, CS}}_T\rvert}{\sigma_Y}$
Q, D non-med.	0.056 (0.012)	0.041 (0.007)	0.072 (0.012)	0.033 (0.003)
D med., Q not	0.059 (0.010)	0.715 (0.073)	0.080 (0.011)	0.091 (0.019)
Q, D med.	0.948 (0.117)	0.354 (0.065)	0.112 (0.014)	0.089 (0.015)

Table 4.2: Results on local mediation analysis.

4.7.1 Robustness to Adversarial Attacks

We empirically show that on-manifold Shapley values are more robust to adversarial attacks than off-manifold Shapley values, as they are computed using a conditional reference distribution. Consequently, PWSHAP effects which marginalise over a conditional reference distribution are also more robust to adversarial attacks than off-manifold Shapley values.

We compare the three explanation methods on a synthetic fairness study. Here, we consider three black-box models: a fair model, an unfair model, and an "attacker model" that returns fair predictions on instances classified as on the data manifold and unfair predictions on instances classified as not belonging to the data manifold. We generate a gender sensitive attribute as $T \sim U(0,1)$, a departmental difficulty indicator $D \sim Bernoulli(\pi(1-T)+(1-\pi)T)$, with $\pi = 0.99$. We also define a continuous test result $Q \sim U(0,1)$. Notably, $D = 1-T$ with high probability, so we take a classifier classifying a given unit (q,d,t) as belonging to the manifold if $t = 1-d$. The fair model is defined $f^{\text{fair}}(t,q,d) = q$, the unfair model as $f^{\text{unfair}}(t,q,d) = \frac{t+td}{2}$, and the attacker model as $f^{\text{attacker}}(t,q,d) = 1_{\{t=1-d\}} f^{\text{unfair}}(t,q,d) + 1_{\{t=d\}} f^{\text{fair}}(t,q,d)$.

Figure 4.3 shows the explanations given by off-manifold Shapley values w.r.t. T, on-manifold Shapley values w.r.t. T, base PWSHAP effects $\Psi_{T\to Y|\emptyset}$, and PWSHAP effets wrt the path $T \to D \to T$ $\Psi_{T\to D\to Y}$ on for the three types of black-box models in Figure 4.3. The on-manifold Shapley values and PWSHAP effects return very similar boxplots for the unfair and attacker models.

Both methods also capture that the unfair model and the attacker model make predictions from the sensitive attribute T while generating a low attribution to T for the fair model. However, note that the boxplot of $\Psi_{T\to D\to Y}$ for the unfair model does not match the theoretical expectation (from Suppl. B.4.3), a limitation that is most likely due to use of a potentially imprecise iterative imputer to fit conditional distributions, combined with division by small weights. Fitting conditional probabilities well to impute missing values remains an active topic of research [117]. For the attacker model, the off-manifold Shapley values are between that for the unfair model and the fair model, as approximately half of the data with columns generated independently is classified as belonging to the manifold and half is not. This illustrates how off-manifold Shapley values are less robust to adversarial attacks than on-manifold Shapley values and derived methods like PWSHAP effects.

4.8 Experiments: Real-World Data

We present a local mediation analysis experiment on the Adult data set from UCI [94], using the causal graph from [82]. Further results and experimental details can be found in the Supplement B.7. The binary outcome denotes whether an individual's income exceeds $50,000 per year. The causal structure of the data is described in the DAG in Figure 4.4. Race was dichotomised into white/non-white. Our individual of interest is a white

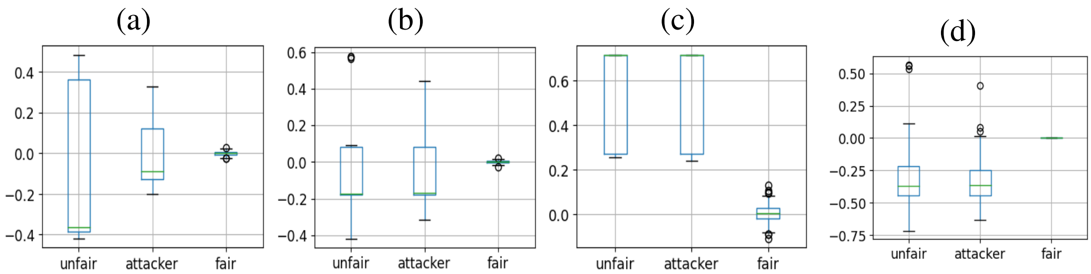


Figure 4.3: Boxplots of (a) off-manifold Shapley values w.r.t. T, (b) on-manifold Shapley values w.r.t. T, (c) base PWSHAP effects $\Psi_{T\to Y|\emptyset}$, (d) PWSHAP effects on the path through D $\Psi_{T\to D\to Y}$, each for the unfair, attacker and fair models.

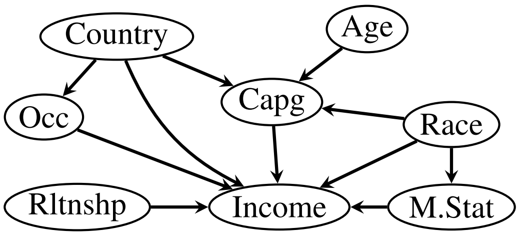


Figure 4.4: DAG of the Census Income Dataset.

38-year-old born in the US, whose marital status (M.Stat) is divorced and Relationship (Rltnshp) is unmarried, and who has a managerial occupation and no capital gain (Capg).

Causal SHAP	ϕ_{direct}	$< 0.001(0.003)$
	ϕ_{indirect}	$0.004(0.006)$
PWSHAP	$\Psi_{\text{Race}\xrightarrow{\text{total}}\text{Inc}}$	$-0.005(0.003)$
	$\Psi_{\text{Race}\to\text{Capg}\to\text{Inc}}$	$0.077(0.004)$
	$\Psi_{\text{Race}\to\text{M.Stat}\to\text{Inc}}$	$0.361(0.004)$

Table 4.3: Results on Census Income Data.

Our aim is to check how (i.e through which paths) being white has influenced our model's prediction of a 0.53 probability of high income for this individual, which is approximately 30 points higher than the average probability in the cohort. Results are shown in Table 4.3, with mean and standard deviation computed by subsampling. PWSHAP shows that the local mediating effect of race through marital status is predominant. Specifying marital status increases the *effect* of race by 0.361, as modelled by our black-box. Causal Shapley results in negligible direct and indirect effects, however we can't readily interpret these quantities obtained from averaging over many coalitions.

This example illustrates three key attributes of PWSHAP compared with Causal Shapley, but also traditional Shapley methods overall: higher resolution, better interpretability of the resulting attributions, true locality of our explanations. The latter characteristic may explain why the effect of moderation by marital status is not captured by Causal Shapley. Here, our individual is an outlier as a large portion of white divorced individuals in the cohort have non-zero capital gain. Both the indirect and direct parts of Causal Shapley are a sum over all coalitions. This implies that for a majority of terms in Causal Shapley race and/or marital status are marginalised over, instead of being set to their feature values. The local effect of race via marital status is thus most likely "blurred" amongst all the coalitions that are "causally redundant" w.r.t. race i.e. where we add/drop features that are independent of race. By contrast, PWSHAP conditions on features that are independent of race (e.g. occupation), which ultimately increases the locality of our method.

4.9 Discussion

PWSHAP shows how Shapley values can generate granular causal explanations for local treatment effects under a posited causal graph. Our results on both simulated and real data show the applicability and locality of our method. The interpretability and safety of PWSHAP make it a strong candidate method for evaluating algorithms in sensitive environments, assuming treatment is binary and a known cause of the outcome.

Practitioners could use PWSHAP explanations to identify inequity or unfairness, prior to deployment. However, our method heavily relies on the cumbersome estimation of the conditional distributions and on the assumed statistical dependencies. Further, dividing by propensity weights can bias the results if the weights are close to zero, and our approach is limited to binary treatments. Following [118], a potential extension of our work may acknowledge the proximity of features in the DAG, instead of only considering features with a direct edge to the target variable. Deriving weights to recover the efficiency property of Shapley is another perspective for future work. PWSHAP could also be used to help guide causal discovery as it can reveal possible conditional independence relationships between variables.

Finally, although PWSHAP is a promising alternative to address rising ethical concerns in AI, note that fairness studies rely on the availability of sensitive data, which can be challenging for practical, ethical or legal reasons. We further caution against relying exclusively on XAI when causal knowledge is insufficient or for black-box without high-accuracy, as misleading interpretations may have

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	PWSHAP: a path-wise explanation model for targeted variables
Publication Status	X Published ◻ Accepted for Publication ◻Submitted for Publication ◻Unpublished and unsubmitted work written in a manuscript style
Publication Details	Ter-Minassian, L.*, Clivio, O.*, Diazordaz, K., Evans, R.J. and Holmes, C.C., 2023, July. PWSHAP: a path-wise explanation model for targeted variables. In International Conference on Machine Learning (pp. 34054-34089). PMLR.

Student Confirmation

Student Name:	Lucile Ter-Minassian		
Contribution to the Paper	I am joint first author of this paper with Oscar Clivio. Under the supervision of Prof. Holmes, I formulated the initial research question: investigating the causal interpretation of the Shapley value of a treatment feature. Oscar and I explored each causal scenario together, with input from Dr. Evans and Dr. Diaz-Ordaz. Oscar primarily derived the lemmas and theoretical properties, while I designed and conducted the experiments. I took the lead in writing the paper, with contributions from Oscar. The non-first authors provided guidance throughout and reviewed the manuscript.		
Signature	LTM	Date	18/02/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments			
Signature	Chris Holmes	Date	25/02/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation

Abstract

Modelling causal effects from observational data for deciding policy actions can benefit from being interpretable and transparent; both due to the high-stakes involved and the inherent lack of ground truth labels to evaluate the accuracy of such models. To date, attempts at transparent causal effect estimation consist of applying post hoc explanation methods to black-box models, which are not interpretable.

Here, we present BICauseTree: an interpretable balancing method that identifies clusters where natural experiments occur locally. Our approach builds on decision trees with a customised objective function to improve balancing and reduce treatment allocation bias. Consequently, it can additionally detect subgroups presenting positivity violations, exclude them, and provide a covariate-based definition of the target population we can infer from and generalize to. We evaluate the method's performance using synthetic and realistic datasets, explore its bias-interpretability tradeoff, and show that it is comparable with existing approaches.

5.1 Introduction

The goal of causal inference is to estimate the strength of a pre-specified intervention on some outcome of interest. It is therefore a valuable tool for data-driven decision-making in public health, economics, political science, and more. These are high-stakes, high-visibility domains in which data-based conclusions should ideally be understood by both the policy makers and the general public, which lack the expertise in the inner workings of statistical models. This highlights the importance of using *interpretable* models for estimating causal effects.

Following [28], *interpretability* means that each decision in the algorithm is inherently explicit and traceable, contrasting with *explainability* where decisions are justified post-hoc using an external model. Therefore, when involving laypeople, interpretable models may often be preferred over explainable ones, as the latter might require understanding the assumptions and implications of an additional explainability model on top of the effect estimation model. Simplicity may incur better trustworthiness.

Interpretable or not, modelling causal effects requires estimating the *potential outcomes* - the outcome a person would experience under every possible intervention [4]. However, at any given time, a person can either be treated or not, but not both. This "fundamental problem of causal inference" [119] implies we never have ground truth labels and thus can never evaluate the accuracy of estimated causal effects. Furthermore, in nonexperimental settings where treatment is not assigned randomly, groups that do or do not receive treatment may not be comparable in their attributes, and such attributes can influence the outcome too, introducing confounding bias. These two difficulties can harm the trustworthiness of data-driven conclusions perceived by both policy makers and the general public. However, by being explicitly transparent, interpretable models allow a new way to evaluate causal models by judging the sensibility of the conclusions they made and whether they adhere to one's common sense.

In this paper, we introduce BICauseTree: Bias-balancing Interpretable Causal Tree, an interpretable balancing method for observational data with binary treatment that can handle high-dimensional datasets. We use a binary decision tree to stratify on imbalanced covariates and identify subpopulations with similar propensities to be treated, when they exist in the data. The resulting clusters act as local naturally randomised experiments. This newly formed partition can be used for effect estimation, as well as propensity score or outcome estimation. Our primary objective is ATE estimation via transparent natural experiment identification. However, our method can further identify positivity-violating regions of the data space, i.e., subsets of the population where treatment allocation is highly unbalanced. By doing so, we generate a transparent, covariate-based definition of the target population, i.e. the population with sufficient overlap for inferring a causal effect.

Our contributions are as follows:

- Our BICauseTree method can identify "natural experiments" i.e. subgroups with lower treatment imbalance, when they exist.
- BICauseTree compares with existing methods for causal effect estimation in terms of bias while maintaining interpretability. BICauseTree is consistent and robust in the sub-populations it creates.
- Our method provides users with a built-in prediction abstention mechanism for covariate spaces lacking common support. We show the value of defining the inferentiable population using a clinical example with matched twins data.
- We show that under mild assumptions (weakly correlated variables, normally distributed data or Gaussian copula model), our framework is guaranteed to decrease the treatment imbalance.
- The resulting tree can further be used for propensity or outcome estimation.
- We release open-source code with detailed documentation for implementation of our method, and reproducibility of our results.

5.2 Related Work

5.2.1 Effect Estimation Methods

Causal inference provides a wide range of methods for effect estimation from data with unbalanced treatment allocation. There are two modelling strategies: *balancing* methods, and *outcome* models.

In *balancing* methods, such as matching or weighting methods, the data is pre-processed to create subgroups with lower treatment imbalance or "natural experiments". *Matching* methods consist of clustering (based on a distance metric) similar units from the treatment and control groups to reduce imbalance. However, as the notion of distance becomes problematic in high dimensional spaces, covariate-based matching tends to become ineffective [120]. *Weighting* methods aim at balancing the covariate distribution across treatment groups, with Inverse Probability Weighting (IPW) [46] being the most popular approach. Sample weights are the inverse of the estimated *propensity* scores, i.e. the probability of a unit to be assigned to its observed group. However, extreme IPW weights can increase the estimation variance.

Contrastingly, *outcome* models estimate the causal effect from regression outcome models where both treatment and covariates act as predictors of the outcome. These regressions can be fitted through various methods like linear regression [7], neural networks [121, 122], or tree-based models [40].

Under this taxonomy, BICauseTree is a *balancing* method, i.e., a data-driven mechanism for achieving conditional exchangeability.

5.2.2 Positivity Violations

Causal inference relies on the *positivity* assumption, which requires covariate overlap between treatment groups. Violations occur when a subgroup rarely or never receives treatment, leading to unreliable causal estimates [123]. Positivity can be handled by estimating propensity scores and trimming extreme values [124], but this reduces interpretability regarding excluded subjects and thus on generalizability. Other methods attempt to improve this by characterizing exclusions or directly assessing covariate overlap without propensity scores [123, 125, 126, 127].

Unlike these approaches, BICause Tree integrates positivity identification into the model, with an interpretable abstention mechanism for effect estimation.

5.2.3 Interpretability and Causal Inference

A critical challenge in many effect estimation methods is their lack of interpretability. A model is *interpretable* when its decisions are inherently transparent, like decision trees where outcomes are easily traced through logical conjunctions. In contrast, *explainable* models rely on post-hoc justifications using explanation tools like Shapley values [57] or LIME [59]. However, these techniques are often unstable and depend heavily on the model's accuracy [128]. Therefore, many practitioners advocate for inherently interpretable models [28], particularly for causal inference, where effect estimation impacts high-stakes decisions involving laypeople.

Beyond effect estimation, the characterization of positivity violations must also be interpretable for stakeholders like policymakers. Excluding samples can reduce external validity, potentially excluding a structurally biased subpopulation. Interpretable identification of the overlap helps policymakers understand to whom the study results apply

[123, 125].

BICauseTree directly generates a covariate-based description of the excluded subgroups, clarifying the target population for which the inferred effect is valid.

5.2.4 Resulting Objectives

Our aim is to bridge the aforementioned gaps in the literature. In turn, our goals are: (i) unbiased estimation of causal effect, (ii) interpretability of both the balancing and positivity violation identification procedures, (iii) ability to handle high-dimensional datasets.

5.3 BICauseTree

Our BICauseTree approach balances observational datasets with the goal of estimating a causal effect in a subpopulation with sufficient overlap. BICauseTree utilizes the Absolute Standardised Mean Difference (ASMD) [129] frequently used for assessing potential confounding bias in observational data. Note that our balancing procedure is entirely interpretable, although it can be used in combination with arbitrary black-box outcome models or propensity score models. Finally, our method generates a covariate-based definition of the target population with reasonable overlap, where our inference is valid.

5.3.1 Formal Causal Model

We consider a dataset of size n. For each individual sample i, we denote (X_i, T_i, Y_i) with $X_i \in \mathbb{R}^d$ a covariate vector for sample i measured prior to binary treatment allocation T_i. In [130]'s potential outcomes framework, $Y_i(1)$ is the outcome under $T_i = 1$, and $Y_i(0)$ is the analogous outcome under $T_i = 0$. Then, assuming the consistency assumption, the observed outcome is defined as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Here, we aim to estimate the average treatment effect (ATE), defined as: $\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$.

5.3.2 Algorithm

The intuition for our algorithm is that, by partitioning the population to maximize treatment allocation heterogeneity, we may be able to find subpopulations that are natural experiments. We recursively partition the data according to the most imbalanced covariate between treatment groups. Using decision trees makes our approach transparent and non-parametric.

Splitting criterion The first step of our algorithm is to split the data until some stopping criterion is met. The tree recursively splits on the covariate that maximizes treatment allocation heterogeneity. To do so, we compute the Absolute Standardised Mean Difference (ASMD) for all covariates and select the covariate with the highest absolute value. The ASMD for X_j is:

$$\mathrm{ASMD_j} = \frac{|\mathbb{E}[X_j|T=1]-\mathbb{E}[X_j|T=0]|}{\sqrt{Var([X_j|T=1])+Var([X_j|T=0])}}$$

Since we assume all confounders are observed, the reason for choosing the one with the highest ASMD is that it is most likely to cause the most confounding bias, and therefore adjusting for it will likely minimize the residual confounding bias (Figures C1 and C2).

Once that next splitting covariate j_{max} is chosen, we want to find a split that is most associated with treatment assignment, so that we may control for the effect of confounding. The tree finds the optimal splitting value by iterating over covariate values $x_{j_{max}}$ and taking the value associated with the lowest p-value from a Fisher's exact test or a χ^2 test, depending on the sample size.

Stopping criterion The tree-building process terminates when either: (i) the maximum ASMD is below a threshold, (ii) the minimum treatment group size falls below a threshold (iii) the total population falls below a threshold, or (iv) a maximum tree depth is reached. All thresholds are user-defined hyperparameters.

Pruning procedure Once the stopping criterion is met in all leaf nodes, the tree is pruned. A multiple hypothesis test correction is first applied on the p-values of all splits. Following this, the splits with significant p-values or with at least one split with significant p-value amongst their descendants are kept. The implementation of the tree allows for user-defined multiple hypothesis test correction, with current experiments using Holm-Bonferonni [131, 132]. The choice of the pruning and stopping criterion hyperparameters will guide the bias/variance trade-off of the tree. Deeper trees may have more power to detect treatment effect while shallower trees will be more likely to have biased effect estimation.

Positivity fiolation filtering Lastly, we evaluate the overlap in the resulting set of leaf nodes to identify those where inference isn't possible. Treatment balance is evaluated using a user-defined overlap estimation method, with the default method being the Crump procedure [37]. The positivity-violating leaf nodes are tagged and excluded from inference.

Estimation Once a tree is contracted, it can be used to estimate both counterfactual outcomes and propensity scores. For each leaf, using the units propagated to that leaf, we can model the counterfactual outcome by taking the average outcome of those units in both treatment groups. Alternatively, we can fit any arbitrary causal model (e.g., IPW or an outcome regression) to obtain the average counterfactual outcomes in that leaf [133, 134]. The ATE is then obtained by averaging the estimation across leaves. Similarly, we can estimate the propensity score in each leaf by taking the treatment prevalence or using any other estimator.

Algorithm 1 BICauseTree

Inputs: root node N_0, X, T, Y
Call *Build subtree*(N_0, X, T, Y)
Do multiple hypothesis test correction on all split p-values
Pruning procedure: keep splits with either (i) a significant p-value or (ii) at least one descendant with a significant p-value
Mark leaf nodes that violate positivity criterion

Algorithm 2 Build subtree

Inputs: current node N, X, T, Y
if Stopping criteria not met **then**
  Find and record in N the covariate with maximum ASMD: $maxASMD := \max_i(ASMD_i)$
  Find and record in N the split value with the lowest p-value according to a Fisher test/χ^2 test
  Record the p-value for this split in N
  Split the data X,T,Y into $X_{left},T_{left},Y_{left}$ and $X_{right},T_{right},Y_{right}$ according to N's splitting covariate and value
  Add two child nodes to N: N_{left} and N_{right}
  Call *Build subtree*(N_{left}, X_{left}, T_{left}, Y_{left})
  Call *Build subtree*(N_{right}, X_{right}, T_{right}, Y_{right})

Decrease in treatment allocation imbalance

Lemma 2 (ASMD Decrease in Node Splitting with Defined Notation). Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of covariates, where each X_i is either continuous or categorical, and let $T \in \{0,1\}$ be a binary treatment indicator. Assume a multivariate normal model or a Gaussian copula model with latent variables $\mathbf{Z} = (Z_1, \ldots, Z_p)$ such that:

- If X_i is continuous, then $X_i = F_i^{-1}(\Phi(Z_i))$, where F_i is the marginal CDF of X_i, and Φ is the standard normal CDF.

- If X_i is categorical, then $X_i = k$ if $Z_i \in (\tau_{i,k-1}, \tau_{i,k}]$ for threshold parameters $\{\tau_{i,k}\}$ shared across treatment groups.

Define:

- ρ_{ij}: correlation between Z_i and Z_j
- σ_i^2: variance of X_i (assumed equal across treatment groups for continuous X_i)
- ASMD_i: absolute standardized mean difference of X_i between treatment groups
- C: constant depending on the truncation point used in the split
- M: upper bound on the ratio of standard deviations of any two continuous variables, i.e., $\frac{\sigma_i}{\sigma_j} \leq M$
- $K < 1$: upper bound on relative ASMD imbalance for other variables compared to X_j
- δ: maximum allowable pairwise correlation, bounded as $\delta \leq \frac{1-K}{CM^2}$

Assume the following conditions hold:

1. For all $i \neq j$, $|\rho_{ij}| \leq \delta \leq \frac{1-K}{CM^2}$
2. For all $i \neq j$, $\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}$
3. For continuous variables, $\frac{\sigma_i}{\sigma_j} \leq M$ for all $i \neq j$
4. For continuous variables, $\text{Var}(X_i \mid T = 1) = \text{Var}(X_i \mid T = 0) = \sigma_i^2$
5. For categorical variables, the threshold parameters $\{\tau_{i,k}\}$ are invariant across treatment groups

Then, splitting the data on the variable X_j with the highest ASMD ensures that the maximum ASMD across all covariates in either resulting child node does not exceed the parent node's maximum ASMD. That is:

$$\max_i \text{ASMD}_i^{\text{child}} \leq \max_i \text{ASMD}_i^{\text{parent}}.$$

This result holds even when variances differ across treatment groups, with modified constants.

Section C.3 provides a proof of convergence.

Code and implementation BICauseTree code is released open-source under: `https://github.com/IBM-HRL-MLHLS/BICause-Trees`. Our flexible implementation allows users to define a stopping criteria as well as a multiple hypothesis correction method. BICauseTree adheres to `causallib`'s API, and can accept various outcome and propensity models.

5.4 Experiment and Results

5.4.1 Experimental Settings

In all experiments–unless stated otherwise–the data was split into a training and testing set with a 50/50 ratio. The training set was used for the construction of the tree and for fitting the outcome models in leaf nodes, if relevant. Causal effects are estimated by taking a weighted average of the local treatment effects in each subpopulation. At the testing phase, the data is propagated through the tree, and potential outcomes are evaluated using the previously fitted leaf outcome model. We performed 50 random train-test splits, which we will refer to as *subsamples*, to avoid confusion with the tree partitions. For each subsample, effects are only computed on the non-violating samples of the population, as defined by BICauseTree. Specifically, we first used BICauseTree to filter overlap-violating observations and then applied all models on that same remaining sample. This ensured i) a fairer comparison, and ii) smaller likelihood of observations with extreme propensity scores.

To maintain a fair comparison, these samples are also excluded from effect estimation with other models and with ground truth.

Baseline comparisons We compare our method to the three most interpretable methods amongst existing approach –double Mahalanobis Matching, Inverse Probability Weighting (IPW), and Causal Tree (CT).

- In **Mahalanobis Matching** [135, 136], the nearest neighbor search operates on the Mahalanobis distance: $d\left(X_i, X_j\right) = \left(X_i - X_j\right)^T \Sigma^{-1} \left(X_i - X_j\right)$, where Σ is the estimated covariance matrix of the control or treatment group.
- In **Inverse Probability Weighting** [46], a propensity score model estimates the individual probability of treatment conditional on the covariates. The data is then weighted by the inverse propensities $P(T = t_i \mid X = \boldsymbol{x}_i)^{-1}$ to generate a balanced pseudo-population.
- In **Causal Tree** [40], the splitting criterion optimizes for treatment *effect* heterogeneity (see section C.4 for further details). We use a Causal Tree and not a Causal Forest

> to compare to an estimator which is equally interpretable as our estimator. Although both BICauseTree and CT utilize decision trees, BICauseTree splits are optimised for balancing treatment *allocation* whilst CT splits are optimised for balancing treatment *effect*, under assumed exchangeability. In other words, CT *assumes* exchangeability while BICauseTree *finds* exchangeability.

As using a single Causal Tree for our interpretability goal gives rise to high estimation bias, Causal Tree was excluded from the main manuscript for scaling purposes. Finally, BiCause Tree also shares some similarities with Coarsedened Exact Matching (CEM) [137], but the strength of our approach lies in performing "blocking" in a principled way, based on ASMD.

We refer the reader to sections C.5.2, C.6 and C.7 for a comparison with CT. For synthetic experiments, we use the simplest version of our tree which we term BICauseTree(Marginal) where the effect is estimated by taking average outcomes in leaf nodes. For real-world experiments, we compare BICauseTree(Marginal) with BICauseTree(IPW), an augmented version in which an IPW model is fitted in each leaf node. To compare estimation methods, we compute the difference between the estimated ATE and the true ATE for each subsample (or train-test partition) and display the resulting distribution of estimation biases in a box plot. Further experimental details, including hyperparameters, can be found in the Appendix under sections C.5 and C.9.

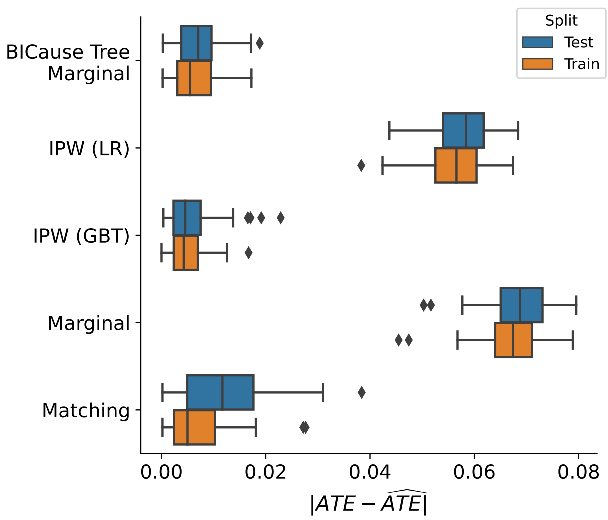


Figure 5.1: Estimation bias for the natural experiment dataset across 50 subsamples, with $N = 20,000$.

5.4.2 Synthetic Datasets

We evaluate BICauseTree on two synthetic datasets: the "natural experiment dataset" to demonstrate its ability to identify subgroups with lower treatment imbalance, and the "positivity violations dataset" to showcase its identification of positivity-violating samples. Due to the interaction-based nature of the data generation procedure, we additionally compare our approach to an IPW estimator with a Gradient Boosting classifier propensity model, referred to as IPW (GBT) in both synthetic experiments. This choice ensures a fair comparison across estimators.

Identifying natural experiments For the natural experiment dataset, we used a Death outcome D, binary treatment T and two covariates:
Sex S and Age A, thus $X = (S, A) \in \mathbb{R}^2$.
We defined four sub-populations, each constituting a natural experiment with a truncated normal propensity distribution centered around a pre-defined mean and variance (see Section C.5.1). Then, individual treatment propensities were sampled from the corresponding distribution, and observed treatment values were sampled from a Bernoulli parameterised with the individual propensities. No positivity violation was modelled. The sample size is $N = 20,000$. The marginal distribution of covariates follows: $S \sim \text{Ber}(0.5), A \sim \mathcal{N}\left(\mu, \sigma^2\right)$ where $\mu = 50, \sigma = 20$.

The partition obtained from training BICauseTree on the entire dataset is shown in Figure C4. Our tree successfully identifies the subpopulations in which a natural experiment was simulated. Figure 5.1 shows the estimation bias across subsamples. In addition to being transparent, BICauseTree has lower bias in causal effect estimation compared to all other methods, excluding IPW(GBT) which has comparable performance. Despite its higher estimation variance, Matching has low bias, probably due to covariate space being well-posed and low-dimensional. Contrastingly, the logistic regression in IPW(LR) is not able to model treatment allocation as the true propensities are generated from a noisy piecewise constant function of the covariates resulting in a threshold effect that explains its poor performance. The non-parametric, local nature of both Matching and BICauseTree thus contrasts with the parametric estimation by IPW(LR).

Identifying positivity violations For the positivity violations dataset, we consider a synthetic dataset with a Death outcome D, a binary treatment of interest T, and three Bernoulli covariates –Sex S, cancer C and arrhythmia A– such that $X = (S, C, A)$ (see Section C.5.1.1 for further details). As for the natural experiment dataset, we modelled treatment allocation with stochasticity by sampling propensities from a truncated Gaussian

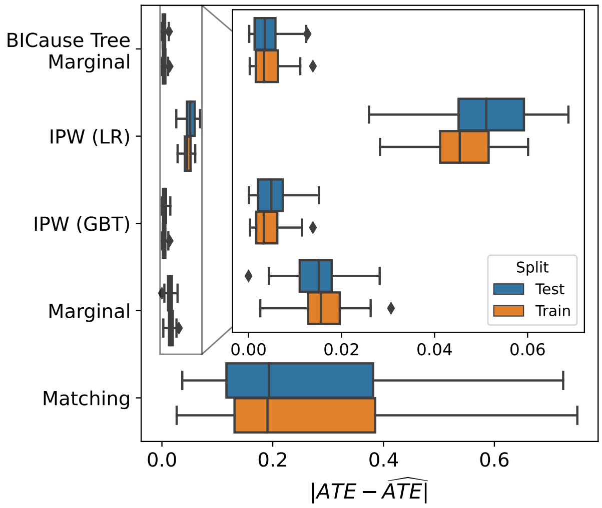


Figure 5.2: Estimation bias across 50 subsamples, positivity violations dataset, excluding positivity violating leaf nodes.

distribution first. Treatment allocation was simulated to ensure that overlap is very limited in two subpopulations: females with no cancer and no arrhythmia are rarely treated, while males with cancer and arrhythmia are almost always treated.

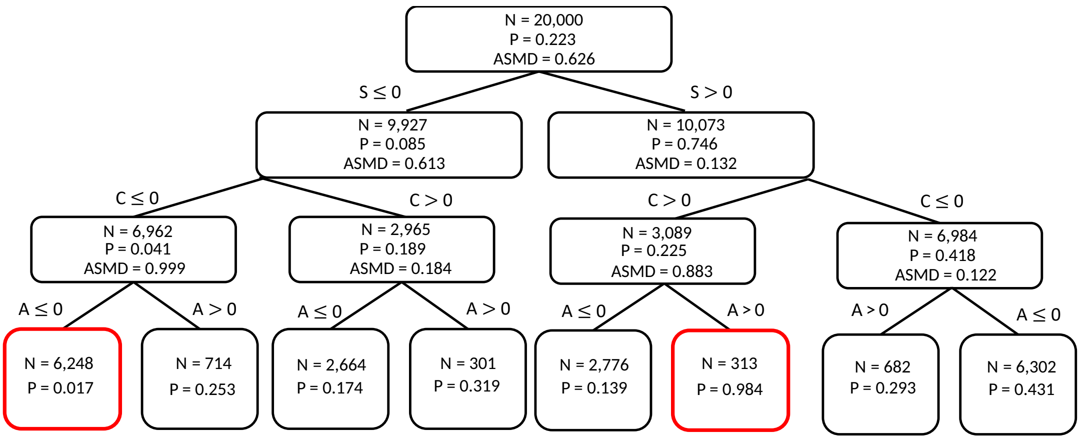


Figure 5.3: Tree structure after training on the entire positivity violations dataset. Violating leaf nodes are in red. P is for treatment prevalence, ASMD is for maximum ASMD.

Figure 5.3 shows the partition obtained from training BICauseTree on the entire dataset, where BICauseTree excludes the subgroups where positivity violations were modelled. This further shows the interpretability of the covariate-based definition of our target population, i.e. the overall cohort, excluding men without cancer and arythmia, and women with cancer and arythmia. On average, 67.1% of the cohort remained after positivity filtering with very little variability across subsamples. Maximum ASMD is null in all leaf nodes.

As seen in Figure 5.2, BICauseTree's effect estimation remains unbiased with low variance after filtering. Our estimator compares favorably with IPW(GBT) while maintaining interpretability. IPW(LR) is more biased, likely due to extreme weights in the initial cohort. Despite filtering samples with overlap violations, the remaining propensity weights may still induce bias. Variance is comparable across methods, except Matching, which is both more biased and has higher variance. Further calibration results for BICauseTree are in the Appendix, section C.5.2.2.

5.4.3 Realistic Datasets

Causal benchmark datasets We use two causal benchmark datasets to show the value of our approach. The twins dataset illustrates the high applicability of our procedure to clinical settings. It is based on real-world records of $N = 11,984$ pairs of same-sex twin births and has 75 covariates. It tests the effect of being born the heavier twin (i.e. the treatment) on death within one year (i.e. the outcome). We use the dataset generated by [138], which simulates an observational study from the initial data by selectively hiding one of the twins with a generative approach. We also ran our analysis on the *2016 Atlantic Causal Inference Conference* (ACIC) semisynthetic dataset with simulated outcomes [139]. For ACIC, given that trees are data greedy, and due to the smaller sample size ($N = 4,802$) relative to the number of covariates ($d = 79$), the models were trained on 70% of the dataset. Further results on the two realistic datasets may be found in sections C.6 and C.7.

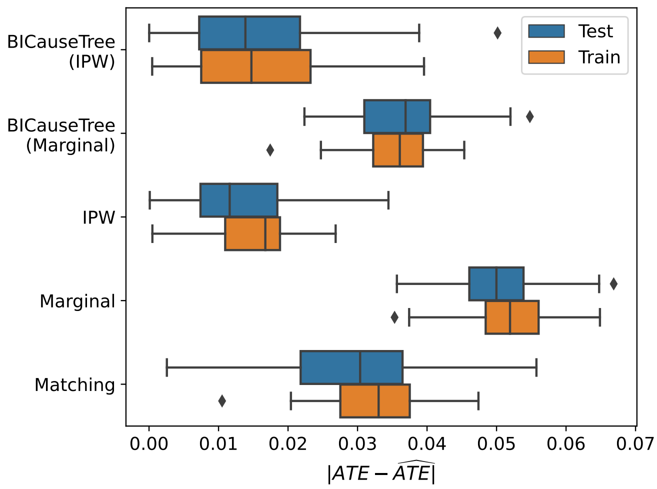


Figure 5.4: Estimation bias for the twins dataset across 50 subsamples, excluding positivity violating leaf nodes.

Effect estimation Figure 5.4 shows the distribution of estimation bias across subsamples on the twins dataset, compared to the baseline models. Here, our BICauseTree(Marginal)

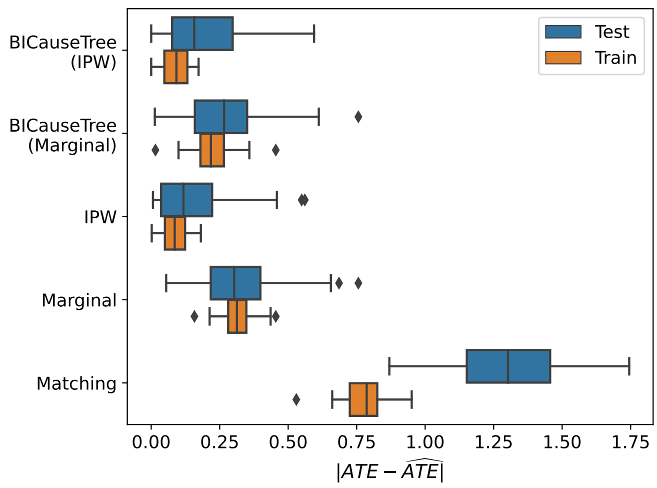


Figure 5.5: Estimation bias for the ACIC dataset across 50 subsamples, excluding positivity-violating leaf nodes.

estimator is less biased than the marginal estimator. Augmenting our tree with an IPW outcome model (BICauseTree(IPW)) further decreases estimation bias, making it comparable with IPW, both w.r.t bias and estimation variance. Figure 5.5 compares the estimation bias on the ACIC dataset. Both BICauseTree models compare with IPW in estimation bias and variance.

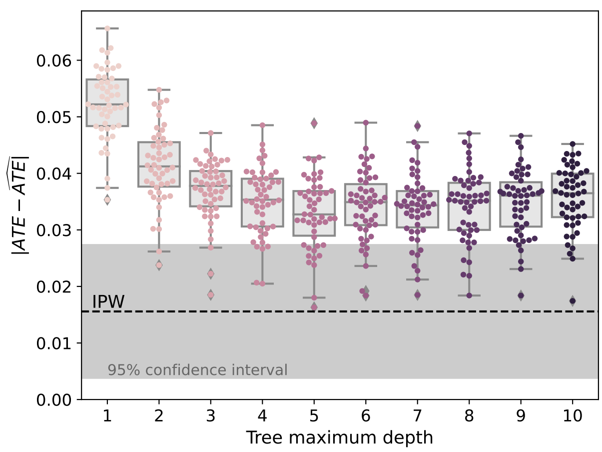


Figure 5.6: Estimation bias for *BICauseTree(Marginal)* with varying maximum depth parameter, and average bias of IPW (dotted), on the twins training set.

Bias-interpretability tradeoff We expect a bias-interpretability tradeoff, where deeper trees are less biased but harder to interpret, while shallower trees are easier to understand but less accurate. Figure 5.6 shows how leaf node bias decreases as we increase the maximum depth of our BICauseTree(Marginal) in the twins dataset (see Figure C23 for

ACIC). Each circle represents bias for a subsample, with the dotted line showing average bias using IPW and the shaded area representing the 95% CI for IPW. The overlap suggests a tradeoff between interpretability and bias, with deeper trees needing more complex outcome models to reduce bias. Augmenting BICauseTree with IPW in the leaves reduces bias but sacrifices transparency.
Ultimately, figure 5.6 shows that bias reduction is consistent beyond a maximum depth parameter of 5. This robustness w.r.t the maximum depth hyperparameter may be due to the pruning step.

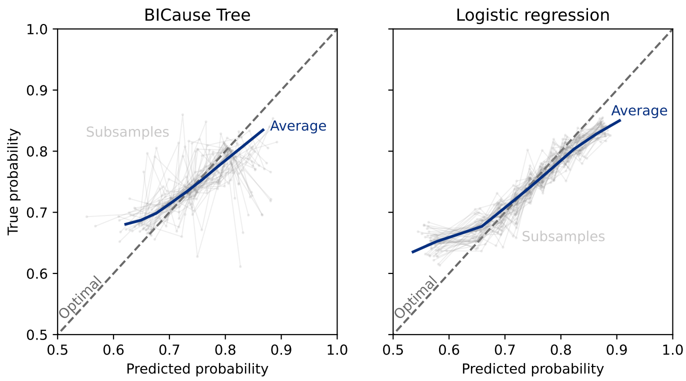


Figure 5.7: Calibration of propensity score, twins dataset

Interpretable positivity violations filtering As previously discussed, BICauseTree provides a built-in method for identifying positivity violations in the covariate space directly. After positivity filtering, the effect was computed on an average of 99.5% ($\sigma = 0.006$) of the population on the twins dataset, and an average of 85.9% ($\sigma = 0.093$) of the ACIC dataset.

Figure C17 shows the tree partition for the twins dataset, identifying one leaf node with positivity violations ($N = 106$). This example highlights the real-world relevance of defining a covariate-based non-violating subpopulation; we can assert that the estimated effect of being born heavier may not apply to newborns in this violating node. BICauseTree's ability to make such distinctions is crucial in safety-sensitive settings. For instance, if the newborns in that violating subgroup are at higher risk, applying the ATE to the entire cohort could deny them necessary follow-up, increasing their risk of death.

Knowing who these newborns are would not be possible using an alternative non-interpretable model that provides opaque exclusion criteria, like propensity trimming. Note that, unlike for effect estimation, the positivity identification remains transparent regardless of the chosen propensity or outcome model at the leaves.

User study for interpretability We further evaluate the interpretability of our method compared to our baselines by conducting interviews. Participants had expertise in Statistics, but no expertise in causal inference. On average, results showed increased intuitiveness and clarity of our method as well as lower time to understand our method compared to others. Details are given in the Appendix under section C.8.

Propensity score estimation BICauseTree can also act as a propensity model, modelling the treatment at its leaf nodes. Given the importance of calibrated propensity scores [140], Figure 5.7 compares the calibration of the propensity score estimation of BICauseTree with the one from logistic regression (IPW) on the testing set of the twins dataset (Figure C22 for ACIC). As expected, logsitic regression, which has better data efficiency, has less-noisy calibration. However, BICauseTree still shows satisfying calibration on average. See the Appendix for propensity score calibration for the synthetic natural experiments and calibration of BICauseTree for the *outcome* variable.

Tree consistency To evaluate the consistency of our clustering across subsamples, we train our tree on 70% of the dataset and compute the adjusted Rand index [141] (further details in section C.2). We chose not to train on 50% of the data here as most of the inconsistency would then be due to the variance between subsamples. For the twins dataset, the Rand index across 50 subsamples of size $N = 8,388$, is 0.633 ($\sigma = 0.208$). For the ACIC dataset, the Rand index across 50 subsamples of size $N = 3,361$, is 0.314 ($\sigma = 0.210$) which shows that our tree is not consistent across subsamples if the sample size is not substantial. However, we exemplify consistent identification of the positivity population, with the variance of the percentage of positive samples being $\sigma = 0.006$ and $\sigma = 0.093$ (see paragraph 5.4.3) in the twins and ACIC dataset respectively. Throughout our experiments, we noticed how consistency starts to decrease if the maximum depth hyperparameter increases past a certain threshold, and thus recommend users test the tree consistency across subsamples when tuning this hyperparameter.

5.5 Discussion

We have introduced a decision tree-based model, with a customised splitting criterion, able to estimate the ATE in a transparent, interpretable way. Additionally, the model can also identify subgroups violating the positivity assumption, characterize them in covariate space, and abstain from estimating an effect on them. Our experimental results on both synthetic data and realistic data show our method has comparable accuracy to existing baselines. Our user study confirms its interpretability is either far superior, or comparable with methods that aren't efficient in high dimensions.

Real-world relevance of our model While our model may retain residual confounding bias, its high interpretability meets the needs of conservative sectors like healthcare, where causal methods face skepticism due to the lack of ground truth [142]. In high-stakes applications like risk assessment and policy decisions, our approach balances performance with transparency, enabling expert scrutiny and fostering trust. Its simplicity also makes it practical for low-resource settings or real-time scenarios, such as generating risk scores for patient prioritization in hospitals based on causal effect measures, even with limited computational resources or time [143, 144].

Overlap filtering Additionally, BICauseTree can detect, characterize, and exclude subgroups violating the overlap assumption, improving internal validity by ensuring outcome extrapolation is data-driven rather than model-based. It also enhances external validity by informing policymakers about the population—explicitly defined in covariate space—to which the estimated effect is expected to generalize. On top of being intuitive, BICauseTree inherits additional desirable strengths from decision trees, like identifying complex interactions in a flexible, data-adaptive, non-parameteric way in high-dimensional data (see Figure C3), with a computational expense comparable to IPW and Causal Forest [145].

Limitations While our method inherits the interpretability of decision trees, it also inherits their weaknesses, such as lower data efficiency compared to parametric estimators like IPW. Theoretically, this leads to higher variance, but since IPW is a low-precision model, our experiments show comparable variance. Additionally, decision trees use fewer samples the deeper they grow, decreasing the precision of the estimated splitting criterion. Therefore, while the mean of the estimated ASMD may be independent of sample size, its variance is dependent and can become less precise [129]. This holds true to our ASMD-based splitting criterion. Furthermore, choosing covariates for split is based solely on ASMD, which, first, is a marginal measure, and second, does not regard the association of covariate and outcome (see C.5.2.1).

Future work Additional research could perfect some of the limitations presented above. First, the pruning procedure correcting for multiple hypotheses can avoid assuming independence and leverage the hierarchical tree structure [146, 147, 148]. Second, following the "honest effect" framework [40] we can further split the sample, using one split to structure the tree and another to fit the node models; though this will require even more data. Third, we currently fit models at each leaf independently, but since our target estimand is the ATE, we can partially pool estimates across leaves (clusters), fitting a multilevel outcome model with varying intercepts or slopes for treatment coefficients [149]. Furthermore, we may learn the tree by measuring the homogeneity at each split (e.g., using the Gini index), thereby defining a loss function to optimize instead of relying on our current greedy search. That loss function could also include a positivity criteria that can be regulated. Lastly, future work may further explore excluding terminal leaf nodes with high imbalance (as they might not enclose natural experiments), or experiment with partially applying additional covariate adjustments in those unbalanced leaves, while balanced uses a simple, interpretable average for estimation.

Limitations notwithstanding, we claim BICauseTree is very relevant when causation is examined in high-stake, public-facing domains. Its transparency can inspire trust among laypeople, while any concerns about its accuracy may be mitigated by fitting an ensemble of causal models and examining if the BICauseTree estimator is an outlier. We provide an open-source implementation of BICauseTree in Python adhering to `causallib`'s API to maximize ease of use and encourage practitioners to adopt it in their work.

Acknowledgements

This work was conducted while LTM was a PhD student at the University of Oxford and an intern at IBM Research Israel. We sincerely thank our reviewers for their valuable feedback, which helped improve this manuscript. Additionally, we are grateful to Michael Danziger for his careful proofreading of this article.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation
Publication Status	◻ Published X Accepted for Publication ◻Submitted for Publication ◻Unpublished and unsubmitted work written in a manuscript style
Publication Details	Ter-Minassian, L., Szlak, L., Karavani, E., Holmes, C. and Shimoni, Y., 2024. Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation. arXiv preprint arXiv:2401.17737. Accepted for publication at the AISTATS 2025 Conference

Student Confirmation

Student Name:	Lucile Ter-Minassian		
Contribution to the Paper	As the lead author, I conceptualized the tree-based formulation under the supervision of Dr. Szlak and Dr. Shimoni. I derived the theoretical result on ASMD reduction and wrote the manuscript. Dr. Szlak and Mr. Karavani contributed to the code implementation and the design of experiments. All non-first authors provided guidance and reviewed the manuscript.		
Signature	LTM	Date	18/02/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments			
Signature	Chris Holmes	Date	25/02/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

Is Merging Worth It ? Securely Evaluating the Information Gain for causal dataset acquisition

Abstract

Merging datasets across institutions is a lengthy and costly procedure, especially when it involves private information. Data hosts may therefore want to prospectively gauge which datasets are most beneficial to merge with, without revealing sensitive information. For causal estimation this is particularly challenging as the value of a merge depends not only on reduction in epistemic uncertainty but also on improvement in overlap.

To address this challenge, we introduce the first *cryptographically secure* information-theoretic approach for quantifying the value of a merge in the context of heterogeneous treatment effect estimation. We do this by evaluating the *Expected Information Gain* (EIG) using multi-party computation to ensure that no raw data is revealed. We further demonstrate that our approach can be combined with differential privacy (DP) to meet arbitrary privacy requirements whilst preserving more accurate computation compared to DP alone. To the best of our knowledge, this work presents the first privacy-preserving method for dataset acquisition tailored to causal estimation. Code is available here.

6.1 Introduction

As the demand for data-driven decision-making grows, the question of how to optimally collect data for a given task becomes increasingly important. Data fusion [150], which integrates pre-existing data from various sources, is a popular method to increase sample size, reduce sampling variability, and enhance statistical power and robustness [151, 152]. However, merging datasets is often a time-consuming and resource-intensive task. This is especially true in sensitive domains such as healthcare, where concerns surrounding privacy, downstream applications, and data security mean long ethical approval procedures are required before undergoing a merge [153, 154, 155]. Consequently, practitioners crucially need methods to determine the value of a potential merge in advance, whilst also complying with privacy requirements [156].

In this work, we focus on data fusion in the context of heterogeneous treatment effect estimation. As a concrete example of this problem, consider a hospital that wishes to assess the impact of a medical intervention on its patients. Upon finding that its own data is insufficient to get accurate estimates, the hospital plans to select one of K possible candidate hospitals for a potential data merge. Given the costs involved, the hospital would like to identify *in advance* which potential dataset would provide the most information upon merging, *whilst* complying with privacy regulations. We propose a solution for such types of problems, allowing for scenarios where patient outcomes at the candidate hospitals to be unobserved or simply masked.

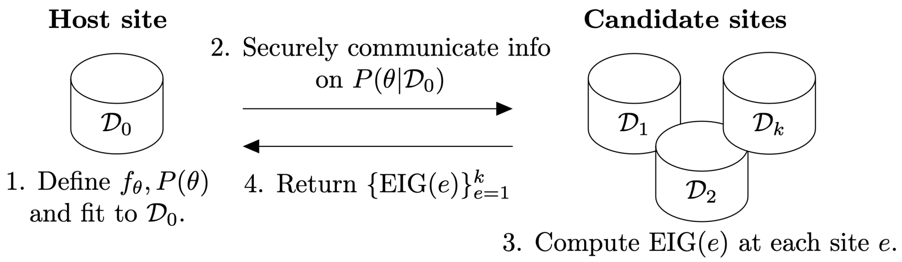


Figure 6.1: Flow chart depicting our method.
In step 1, the *host site* chooses a parameterised model class for the conditional outcome, given by f_θ and sets the prior $P(\theta)$. They then perform a local Bayesian update, and communicate information on the posterior $P(\theta|\mathscr{D}_0)$ to the *candidate site* using secure multi-party computation (MPC). The candidate applies MPC once more to privately calculate the Expected Information Gain (EIG) and communicate it back to host. This allows the host to select the best merge out of the potential candidates.

We quantify the value of a merge in a principled, information theoretic way by applying techniques from Bayesian experimental design [157, 158, 159]. Our solution shares similarities with standard Bayesian dataset acquisition [160, 161], however in causal contexts data serve an additional, distinct purpose. Specifically acquired data should not only enhance our understanding of the outcome function, but also assist in combating the selection bias that is inherent to causal effect estimation [119] by balancing treatment. Put differently, whilst generic data fusion aims at reducing the epistemic uncertainty from incomplete knowledge of the outcome, causal estimation also seeks to improve treatment overlap [37, 162], i.e. reducing the epistemic uncertainty for counterfactual outcomes. To resolve this, we make use of favourable parameterisations available in popular Bayesian causal inference methods [1, 163], which allows us to prioritise information gain in the parameters relevant for the causal problem, rather than gaining information in irrelevant parts of the conditional outcome.

To ensure privacy, we employ Secure Multi-Party Computation [164, 165, 166]. This cryptographic protocol enables multiple parties to jointly compute the output of a function without revealing any of their own private inputs. A classic example involves determining the wealthiest person in a group without anyone revealing their personal net worth. In our context, it allows the different candidate sites to compute their expected information gains relative to the initial sites' data, without exposing the contents of their datasets. This ensures that the noise required for privacy guarantees can be added to the final statistic as opposed to the underlying data at each site.

Our contributions can be summarised as follows:

- We propose information-theoretic methods to measure the value of a data merge in the context of heterogeneous treatment effects estimation. Our primary contribution is a novel approach that specifically targets the reduction of entropy in parameters that directly influence the conditional average treatment effect (CATE). Additionally, we also present a standard approach based on expected entropy reduction in all parameters of a conditional outcome model.
- We demonstrate how both of the approaches can be used with three popular CATE estimators; Bayesian Polynomial Regression [167], Causal Multitask Gaussian Processes [1], and Bayesian Causal Forests [163]. We derive closed form expressions for the expected entropy reduction in the first two models and give a Monte Carlo estimator for the other.
- We provide a privacy protocol for our methods based on multi-party computation [MPC; 164]. This ensures that statistic can be computed without any party revealing

their raw data. Therefore, differential privacy [DP; 168] guarantees can be achieved by noising the *final* computed statistic, rather than to the original raw data, ensuring less loss of accuracy.

- We experimentally validate our methodology across a range of synthetic and semi-synthetic tasks, demonstrating strong agreement between our prospective rankings and the true rankings obtained after performing the merge. Moreover, we find that our proposed methodology to target CATE parameters improves over traditional Bayesian data selection and a number of other baselines. Finally, we show that, for the same level of privacy guarantees, our MPC protocol chained with DP performs better than applying DP to the raw inputs in the linear case.

6.2 Problem Statement, Assumptions & Notation

Notation The random variables X, T, and Y represent the covariates, treatment, and outcomes, with domains $\mathscr{X}$, $\{0,1\}$, and $\mathscr{Y}$, respectively; $\mathbf{x}$, t, and y denote realisations of these variables. We let $\mathscr{D}_e$ be the dataset comprised of elements $\{\mathbf{x_i}, t_i, y_i\}_{i=1}^{n_e}$, drawn i.i.d. from a distribution $P_e(\mathbf{x}, t, y)$, where $e \in \{0, \cdots, K\}$ indexes the datasets. Vectors of observations in $\mathscr{D}_e$ are denoted in bold, i.e. $\mathbf{y_e} = (y_i)_{i=1}^{n_e}$, $\mathbf{t}_e = (t_i)_{i=1}^{n_e}$, and $\mathbf{X}_e = (\{x_i\})_{i=1}^{n_e}$ refers to the data matrix. We use potential outcomes framework [4], so that $Y(t)$ represents the outcome resulting from an intervention setting $T = t$.

Assumptions and Objective Throughout we focus on estimating the *Conditional Average Treatment Effect* (CATE), given by:

$$\tau(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) | X = \mathbf{x}]$$

To estimate CATE, we begin with an initial dataset, $\mathscr{D}_0$, referred to as the *host*. Our goal is to accurately estimate CATE with respect to the distribution of this dataset, $P_0(\mathbf{x})$. Specifically, we focus on minimising the Precision in Estimation of Heterogeneous Effects [PEHE; 169] given by $\varepsilon_{\text{PEHE}}(f) = \int_{P_0(\mathbf{x})} \left(\hat{\tau}_f(\mathbf{x}) - \tau(\mathbf{x})\right)^2 d\mathbf{x}$, where $\hat{\tau}_f$ is the CATE estimate arising the outcome model f.

We consider a set of potential datasets for merging, $\mathscr{D}_e$, referred to as the *candidate* sites. In these candidate datasets, we assume that the outcomes are unmeasured or masked, and so denote them by $\mathscr{D}_e = \{\mathbf{x}_i^e, t_i^e, Y_i^e\}_{i=1}^{n_e}$ to show the randomness in Y_i^e. The goal is to prospectively identify which of the candidate datasets $\mathscr{D}_e$, would reduce the uncertainty over CATE if we were to measure or unmask the Y_i^e's and merge with the host dataset $\mathscr{D}_0$.

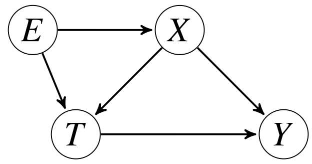


Figure 6.2: Assumed DAG

We assume that datasets are generated according to the Directed Acyclic Graph (DAG) in Figure 6.2.[1] This implies that $P(y|\mathbf{x},t,e)$ is fixed across environments, but both covariate distributions and treatment allocation schemes are free to vary i.e. $P(\mathbf{x},t|e)$ can depend on e. We also assume positivity over the whole population, so that $\forall\ \mathbf{x}, 0 < P(T=1|X=\mathbf{x}) < 1$, but allow for violations at the site level. Adding the consistency assumption (i.e. $Y(t)=Y$ when $T=t$) to positivity and the causal structure in Figure 6.2 (which implies no hidden confounders) we have that CATE is identifiable [170, 171], constant across environments e, and given by:

$$
\begin{aligned}
\tau(\mathbf{x}) &= \mathbb{E}\left[Y|X=\mathbf{x},T=1\right] - \mathbb{E}\left[Y|X=\mathbf{x},T=0\right] \\
&= \mathbb{E}\left[Y|X=\mathbf{x},T=1,E=e\right] \\
&- \mathbb{E}\left[Y|X=\mathbf{x},T=0,E=e\right]
\end{aligned}
$$

6.3 Method

We take a Bayesian approach to modelling the CATE. Specifically, we define $f_\theta : \mathscr{X} \times \{0,1\} \to \mathscr{Y}$ to be a function representing the expected outcome conditional on covariates and treatment, i.e., $f_\theta(\mathbf{x},t)$ seeks to approximate $\mathbb{E}\left[Y|X=\mathbf{x},T=t\right]$. We assign a prior distribution $P(\theta)$, to the parameters θ and denote the conditional likelihood of the outcome given parameters, covariates, and treatment as $P(Y|\theta,\mathbf{x},t)$, which is chosen appropriately based on the type of outcome being modelled.
For example, we can select a normal likelihood for continuous outcomes, or a Bernoulli likelihood for binary ones. The data-generating process can be written as:

$$
\theta \sim P(\theta), \quad Y|f_\theta(\mathbf{x},t) \sim P(Y|\theta,(\mathbf{x},t)).
$$

Throughout this paper we focus on continuous y and use a normal likelihood with fixed variance σ^2, i.e. $P(Y|\mathbf{x},t,\theta) = \mathscr{N}(Y; f_\theta(\mathbf{x},t),\sigma^2)$. CATE is then estimated via the current posterior mean, so that if we have conditioned on data $\mathscr{D}$ our estimate is $\hat{\tau}(\mathbf{x}) = \mathbb{E}_{\theta\sim p(\theta|\mathscr{D})}[f_\theta(\mathbf{x},1) - f_\theta(\mathbf{x},0)]$.

[1] We make us of the SWIG framework to combine causal graphical models with potential outcomes. More details can be found in Richardson and Robins [170].

6.3.1 Quantifying Data Merge Utility through Expected Information Gain

In order to quantify the value of a data merge, we draw inspiration from Bayesian experimental design [BED; 157, 159]. BED applies information theory to provide a measure of what performing a particular experiment would tell us about a parameter of interest, relative to our current beliefs about the parameter value. In our context, the 'design' of the experiment corresponds to choosing a dataset $\mathscr{D}_e$, and the outcome is observing $\{Y_i^e\}_{i=1}^{n_e}$. The value of an experiment is quantified via the expected information gain [EIG; 158], which measures the expected reduction in uncertainty in the parameters, as measured by Shannon entropy, when moving from the post host posterior, $P(\theta|\mathscr{D}_0)$ to the post merge posterior, $P(\theta|\mathscr{D}_0, \mathscr{D}_e)$:

$$\mathrm{EIG}_{\theta|\mathscr{D}_0}(e) = \mathbb{E}[H[P(\theta|\mathscr{D}_0)] - H[P(\theta|\mathscr{D}_0, \mathscr{D}_e)]],$$

where the expectation is over $P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0) = \mathbb{E}_{P(\theta|\mathscr{D}_0)}[P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)]$—the Bayesian marginal distribution of $\mathbf{y}_e$.

The EIG is equivalent to the mutual information between parameters and outcomes under the design and can be equivalently written as:

$$\mathrm{EIG}_{\theta|\mathbf{D}_0}(e) = \mathbb{E}\left[\log \frac{P(\mathbf{y}_e|\theta, \mathbf{X}_e, \mathbf{t}_e)}{P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0)}\right], \tag{6.1}$$

where the expectation is over $P(\mathbf{y}_e, \theta|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0)$.
This form known as Bayesian Active Learning by Disagreement (BALD) in the active learning literature [172]. For certain classes of models, such as polynomial regression or Gaussian processes, the EIG is available in closed form [173]. In other cases if the likelihood function isn't analytically available, we can approximate it using nested Monte Carlo [NMC; 174, 175] as follows:

$$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{P(\mathbf{y}_e^{(i)}|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0)}, \tag{6.2}$$

where $\theta^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta|\mathscr{D}_0)P(\mathbf{y}_e|\theta^{(i)}, \mathbf{X}_e, \mathbf{t}_e)$, and further M_1 samples $\theta'_j \sim P(\theta|\mathscr{D}_0)$ for the denominator

$$\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0) = \frac{1}{M_1}\sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j, \mathbf{X}_e, \mathbf{t}_e\right). \tag{6.3}$$

to give:

$$\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathscr{D}_0}(e) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{1}{M_1}\sum_{j=1}^{M_1}P(\mathbf{y}_e^{(i)}|\mathbf{X}_e,\mathbf{t}_e,\theta'_j)\right)$$
$$-\frac{n_e}{2}(1+\log(2\pi\sigma^2)).$$

Nested estimators are biased but consistent, converging to the true EIG at a rate $\mathscr{O}((N^{-1}+cM_1^{-2})^{\frac{1}{2}})$ [174]. A detailed algorithm for both estimators is given in the Appendix under section D.1.1.

These estimators allow us to gauge the value of a merge before observing outcomes $\{Y_i^e\}_{i=1}^{n_e}$ in the dataset $\mathscr{D}_e$. However, whilst this approach provides us with information on the parameters for the conditional outcome, it may not necessarily lead to improved CATE predictions. This is because $\mathrm{EIG}_{\theta|\mathscr{D}^-}$, encourages *uniform entropy reduction* across all dimensions of θ, not just the ones relevant for CATE estimation. For instance, a dataset with untreated individuals only would provide information about the conditional outcome, but less for CATE, which requires viewing treated individuals as well. This motivates our improved methodology, which we present in the following section.

6.3.2 EIG Targeting CATE Parameters

Many causal inference models have additional parameter structure that allows us to target CATE estimation more directly. Specifically, the parameter set, θ can often be split as $\theta = \theta_{\mathrm{c}} \cup \theta_{\mathrm{nc}}$ where θ_{c} parameterises the CATE model and θ_{nc} is a set of nuisance parameters. For example, in Bayesian Causal Forests [163] the model is parameterised as $f_\theta(\mathbf{x},t) = \mu_{\theta_{\mathrm{nc}}}(\mathbf{x}) + t\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$, where $\mu_{\theta_{\mathrm{nc}}}$ and $\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$ jointly model the conditional outcome, and $\tau_{\theta_{\mathrm{c}}}(\mathbf{x})$ directly models CATE.

Leveraging such a parameterisation, we can prioritise uncertainty reduction in θ_c, and therefore CATE, by maximising:

$$\mathrm{EIG}_{\theta_c|\mathscr{D}_0}(e) = \mathbb{E}\left[\log\frac{P(\mathbf{y}_e|\theta_{\mathrm{c}},\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}{P(\mathbf{y}_e|\mathbf{X}_e,\mathbf{t}_e,\mathbf{D}_0)}\right], \tag{6.4}$$

where expectation is over $P(\mathbf{y}_e,\theta_{\mathrm{c}}|\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)$. Here $P(\mathbf{y}_e|\theta_c,\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0) = \mathbb{E}_{P(\theta_{nc}|\theta_c,\mathscr{D}_0)}[P(\mathbf{y}_e|\theta,\mathbf{X}_e,\mathbf{t}_e)]$ is the Bayesian distribution of the outcomes conditional on the CATE-related parameters only, and is generally not available in closed form. To deal with this intractability, we suggest approximating both the numerator and denominator empirically, yielding the following estimator:

$$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta_c|\mathscr{D}_0}(e) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)},\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}, \tag{6.5}$$

where $\theta_c^{(i)}, \mathbf{y}_e^{(i)} \sim P(\theta_c|\mathscr{D}_0)P(\mathbf{y}_e|\theta_c, \mathbf{X}_e, \mathbf{t}_e)$, the denominator $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0)$ is as in Equation 6.3, and use further M_2 samples $\theta_{nc}^{(ik)} \sim P(\theta_{nc}^{(ik)}|\theta_c^{(i)}, \mathscr{D}_0)$ for the numerator:

$$\widehat{P}(\mathbf{y}_e^{(i)}|\theta_\mathrm{c}^{(i)}, \mathbf{X}_e, \mathbf{t}_e, \mathscr{D}_0) = \frac{1}{M_2}\sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)}, \mathbf{X}_e, \mathbf{t}_e\right)$$

This ensures that we prioritise a gain in information in the part of the model directly responsible for CATE. We again give an algorithm in the Appendix under section D.1.1.

Relation between EIG increase and precision of CATE In the context of conditional average treatment effect (CATE) estimation, the expected information gain (EIG) quantifies the anticipated reduction in posterior uncertainty about heterogeneous effects afforded by new data. An increase in EIG corresponds to a steeper decrease in the posterior variance of the CATE function, thereby enhancing the precision of heterogeneous effect estimates. This relationship emerges from Bayesian decision theory [176, 177]: by prioritizing experiments or sampling strategies that maximize EIG, one effectively selects information-rich observations that sharpen the posterior distribution of individual treatment effects. Consequently, higher EIG not only accelerates convergence of the estimated CATE towards its true value but also yields tighter credible intervals, reflecting improved confidence in the detection and quantification of treatment effect heterogeneity.

6.3.3 Procedure and Model Classes

We now apply both procedures to three popular Bayesian causal inference methods: Bayesian Polynomial Regression, Bayesian Causal Forests [163], and Causal Multi-task Gaussian Processes [1]. We describe the standard predictive method as well as the parameter split used to target CATE.

Bayesian Polynomial Regression Due to its ubiquity across numerous fields, we first apply our method to Bayesian Polynomial Regression model. This involves specifying an initial polynomial transformation[2] $\phi : \mathscr{X} \times \mathscr{T} \to \mathbb{R}^p$, a mean μ_0, and precision matrix Λ_0 to parameterise the prior as:

$$f_\theta(\mathbf{x}, t) = \phi(\mathbf{x}, t)^\top \theta, \quad \theta \sim \mathscr{N}(\mu_0, \sigma^2 \Lambda_0^{-1}). \tag{6.6}$$

ϕ allows for higher order, or interaction terms. We further assume $\phi(\mathbf{x}, t)$ and θ can be split as:

$$\phi(\mathbf{x}, t) = \begin{bmatrix}\phi_\mathrm{nc}(\mathbf{x}) & t\phi_\mathrm{c}(\mathbf{x})\end{bmatrix}^\top, \quad \theta = \begin{bmatrix}\theta_\mathrm{nc} & \theta_\mathrm{c}\end{bmatrix}^\top$$

[2]This could be an arbitrary non-linear function but we focus on polynomial transformations.

This covers a broad range of Bayesian polynomial regressions, including the selection used in Gelman et al. [167]. In the Appendix, section D.2.2.1 shows that both EIGs can be computed in closed form:

Proposition 3. *For the Bayesian Polynomial Regression model defined in Equation 6.6 we have:*

$$\begin{aligned}\mathrm{EIG}_{\theta|\mathscr{D}_0}(e) &= \log\det\left(\Phi_e^\top \Phi_e + \Phi_0^\top \Phi_0 + \Lambda_0\right) + C\\ \mathrm{EIG}_{\theta_\mathrm{c}|\mathscr{D}_0}(e) &= \log\det(\Phi_{c,e}^\top \Phi_{c,e} + \Phi_{c,0}^\top \Phi_{c,0} + \Lambda_0^{[\mathrm{c,c}]}) + C',\end{aligned}$$

where $\Phi_e = \phi(\mathbf{X}_e, \mathbf{t}_e)$*,* $\Phi_0 = \phi(\mathbf{X}_0, \mathbf{t}_0)$*,* $\Phi_{c,e} = \mathbf{t}_e \odot \phi_\mathrm{c}(\mathbf{X}_e)$*,* $\Phi_{c,0} = \mathbf{t}_0 \odot \phi_\mathrm{c}(\mathbf{X}_0)$*, with* ϕ, ϕ_c *applied row-wise and* $\odot$ *denoting element-wise multiplication;* C, C' *are constant in e.*

Bayesian Causal Forest Bayesian Causal Forests [BCF; 163] are one of the most popular causal inference methods, building upon Bayesian Additive Regression Trees [BART; 178], which are themselves a mainstay in observational causal inference [179]. The BCF model can be expressed as $f_\theta(\mathbf{x}, t) = \mu_{\theta_\mathrm{nc}}(\mathbf{x}) + t\tau_{\theta_\mathrm{c}}(\mathbf{x})$,
where μ_{θ_nc} and $\tau_{\theta_\mathrm{c}}(\mathbf{x})$ are independent BART models; further details and alternative parameterisations can be found in the Appendix under section D.2.1. As the posterior is only available via sampling, we estimate both EIGs using NMC as given in Equation 6.2 and Equation 6.5.

Causal Multi-task Gaussian Processes Causal multi-task Gaussian Processes [1] use a vector-valued GP [180] to jointly model the conditional outcomes, allowing information sharing between them:

$$\mathbf{f} = \begin{bmatrix} f(\mathbf{x}, 0) & f(\mathbf{x}, 1) \end{bmatrix}^\top, \text{ where } \mathbf{f} \sim \mathscr{GP}(0, \mathbf{K}) \tag{6.7}$$

for a vector-valued kernel $\mathbf{K} : \mathscr{X} \times \mathscr{X} \to \mathbb{R}^{2\times 2}$. The outcomes are then modelled by evaluating the relevant portion of the GP. Under this setting CATE is given by $\tilde{\tau} = \mathbf{ef}$ for $\mathbf{e} = \begin{bmatrix} -1 & 1 \end{bmatrix}$, meaning that $\tilde{\tau}$ is also a GP given by $\tilde{\tau} \sim \mathscr{GP}(0, \mathbf{e}^\top \mathbf{K} \mathbf{e})$. The advantage of causal multi-task GPs is that they allow us to get a closed form posterior for τ without having to observe any samples from $Y(1) - Y(0)$.

As GPs are inherently non-parametric they do not fit directly into the framework laid out in Sections 6.3.1 and 6.3.2. This is not a problem for the predictive case as we can replace θ with $\mathbf{f}$ in Equation 6.1 and get closed form expressions [172]. However, for the causal case this creates challenges as we cannot directly evaluate the expressions in Equation 6.4 with $\tilde{\tau}$ in the place of θ_c. To resolve this we instead focus on entropy reduction in CATE predictions on the host dataset. We denote this by $\tilde{\tau}(\mathbf{X}_0)$, where $\mathbf{X}_0$ is the host data matrix. The information gains e denote these by $\mathrm{EIG}_{\mathbf{f}|\mathscr{D}_0}(e)$ and $\mathrm{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathscr{D}_0}(e)$ respectively. As the following proposition shows, both of these are now available in closed form:

Proposition 4. *Let $n_e^{(t)}$ be the number of subjects receiving treatment t in dataset e. For the causal multi-task GP model, defined in Equation 6.7 we have*

$$\mathrm{EIG}_{\mathbf{f}|\mathscr{D}_0} = \frac{1}{2}\log\det(\Sigma_1) - n_e^{(0)}\log(\sigma_0) - n_e^{(1)}\log(\sigma_1)$$
$$\mathrm{EIG}_{\tilde{\tau}(\mathbf{X}_0)|\mathscr{D}_0}(e) = \frac{1}{2}\log\left(\det(\Sigma_1)\det(\Sigma_2)\right) - \frac{1}{2}\log\left(\det(\Sigma)\right),$$

where $\Sigma_1, \Sigma_2, \Sigma$ and the proof are given in the Appendix under section D.2.3.

6.4 Privacy

For privacy we use *Multi-Party Computation* [MPC; 165]. First introduced by Yao [164], MPC focuses on a setting where m separate parties wish to compute the value of a function $f(x_1,\cdots,x_m)$ where the i^{th} party inputs x_i and wishes to keep this private. To resolve this, MPC involves the specification of a protocol of message passing between parties which if followed would lead to the computation of f. In this work, we focus on the *semi-honest* setting, in which all parties follow the specified protocol, but some *corrupt* parties will try to learn as much about their peers inputs in the process. The goal is to devise a protocol which will preserve the privacy of the non-corrupt party's inputs, up to a given computational budget by the adversary. In our setting this means that any collection of corrupt sites are unable to learn anything about the other sites data during the EIG calculation, so any noise needed for privacy guarantees can be added to the final statistic.

For implementing multi-party computation, we employ the open source library CrypTen [166]. CrypTen builds upon PyTorch [181] allowing for standard tensor operations to be performed in an MPC protocol. For arithmetic operations on floating-point values this is achieved as follows: A float, x_F, is multiplied by some large scaling factor $B = 2^L$ and rounded to the nearest integer $\lfloor x_F \rceil$, where L is the number of precision bits. The integer $\lfloor x_F \rceil$ is then associated with its equivalence class $x \in \mathbb{Z}/Q\mathbb{Z}$ where $\mathbb{Z}/Q\mathbb{Z}$ is a ring of Q elements. The value x can then be shared across all m parties using Shamir secret sharing [182], in which each party gets access to a share of x is given by $[x]_i \in \mathbb{Z}/Q\mathbb{Z}$ which is generated such that the sum of all shares recovers the original value, so $x = \sum_{i=1}^m [x]_i \mod Q$. At any point all parties can combine their shares to decode the output as $x_F \approx x/B$. We let $[x] = \{[x]_i\}_{i=1}^m$ denote the set of all shares corresponding to the secret value x.

Arithmetic operations building on addition are performed locally, so that for two secret values $[x], [y]$ each party performs $[z]_i = [x]_i + [y]_i$ and the result, z is obtained by all parties summing their share. Multiplication is implemented using Beaver triples [183], logarithms are approximated using householder iterations [184], and reciprocals use Newton-Raphson. We implement log-determinants using Cholesky LDL decompositions, which are preferred to standard Cholesky factorisations as they avoid the use of square roots, which would require additional approximation in Crypten. This is possible as we only compute the log determinant of positive semi-definite matrices. This provides all the operations necessary to implement the above EIG calculations in a private manner using MPC.

When returning EIG statistics to the host, we add a small amount of noise to prevent information leakage. If we only need to output the best site, we use the exponential protocol [168] ensuring minimal information leakage. Finally, we note that the discretisation required to represent a float in the ring $\mathbb{Z}/Q\mathbb{Z}$ involves some degree of precision loss. Nevertheless, as we empirically demonstrate in the following section, this leads to minimal depreciation in performance compared to differential privacy.

6.5 Experiments & Results

We experimentally validate our approach in a setting where the host has to rank a number of candidate datasets based on the estimated gain from merging. We use selection of synthetic and semi-synthetic benchmark datasets as this allows us to use the known CATE to get a ground truth ranking. For each model, we do this by ranking datasets on the true PEHE on $P_0(\mathbf{x})$ of the relevant model trained on the merged dataset $\mathscr{D}_0 \cup \mathscr{D}_e$. This is then compared against implied EIG rankings. Throughout, we tune any model hyper-parameters on the host site when measuring the information gain as well as re-tuning paramters on each merged dataset when getting the ground truth ranking.

To generate the datasets, we begin with an initial, large dataset $\mathscr{D}$ from which we subsample the host and candidate sites. We do this by choosing a selection function, $S_e(\mathbf{x}, t)$, and using it to subsample a dataset of size n_e, where $S_e(\mathbf{x}_i, t_i)$ is the probability of subsampling point i for dataset $\mathscr{D}_e$. Varying the selection functions across sites ensures heterogeneity amongst datasets whilst complying with the independences given by the causal structure in Figure 6.2. We begin by providing an illustrative example on synthetic data before evaluating on the causal benchmarks, specifically: Lalonde [185] and Infant Health and Development Program [IHDP; 169]. Further details and results can be found in the Appendix under section D.3 and section D.4, respectively.

6.5.1 Illustrative Experiment

Concept To illustrate the difference between our two methods—the standard approach based on the full parameter set, EIG_θ (Equation 6.1), and the CATE-targeting one, EIG_{θ_c} (Equation 6.4)—we start with a simple example where the host must choose between two candidate sites. We design these sites as follows: a *complementary* site, representing the "ideal" merge for causal purposes, and a *twin* site, containing information similar to the host.

To create these, we first simulate an initial large dataset $\mathscr{D}$ as if it were a randomised controlled trial with an equal probability of treatment and subsample the host dataset, $\mathscr{D}_0$, using a selection function $S_0(\mathbf{x},t)$. The data of the complementary site, $\mathscr{D}_{comp}$, is subsampled using $1 - S_0(\mathbf{x},t)$ as a selection function, whilst the twin site uses $S_0(\mathbf{x},t)$. This ensures the twin dataset mirrors the host's distribution and the complementary causally "complements" it.

Assuming equal sizes, merging $\mathscr{D}_0$ with $\mathscr{D}_{comp}$ would recreate the initial randomised trial $\mathscr{D}$, as the variable e acts as a collider for $\mathbf{x}$ and t, removing their conditional dependency (see causal DAG in Figure D1 in the Appendix). Therefore, $\mathscr{D}_{comp}$ represents an ideal merge as it balances treatment allocation, whereas $\mathscr{D}_{twin}$ covers similar regions of the data space to those in the host ($\mathscr{D}_0$), potentially amplifying pre-existing biases and imbalances.

For the illustrative experiment, we vary the ratio of sample sizes, $\frac{n_{twin}}{n_{comp}}$, in order to compare which dataset is chosen by $\text{EIG}_{\theta|\mathscr{D}_0}$ and the causally targeted $\text{EIG}_{\theta_c|\mathscr{D}_0}$ for linear regression. The aim is to demonstrate that $\text{EIG}_{\theta|\mathscr{D}_0}$ will prefer $\mathscr{D}_{twin}$ at points where $\mathscr{D}_{comp}$ dataset is still the preferable dataset for CATE estimation. Indeed, for large values of the ratio $\frac{n_{twin}}{n_{comp}}$, $\mathscr{D}_{twin}$ provides significant information about the conditional outcome, but not in the regions that are most relevant for causal estimation. On the other hand, EIG_{θ_c} should continue to select $\mathscr{D}_{twin}$ whilst it remains preferable for CATE estimation. We simulate $\mathbf{x} \in \mathbb{R}^3$ where x_1 is Bernoulli and other covariates are normal. The true outcome is sampled from a normal linear model, and the selection functions are logistic regressions. We also include sample based estimates for each EIG to demonstrate how they differ from their closed form counterparts. Experimental details provided in section D.3 in the Appendix.

Results Figure 6.3 shows the results of the experiment divided into three regions. In region one, both methods choose the complementary dataset over the twin dataset which is consistent with ground truth ranking given by the PEHE upon merging. In region two, $\text{EIG}_{\theta|\mathscr{D}_0}$ chooses $\mathscr{D}_{twin}$ whilst $\text{EIG}_{\theta_c|\mathscr{D}_0}$ opts for $\mathscr{D}_{comp}$. Here, the complementary dataset is

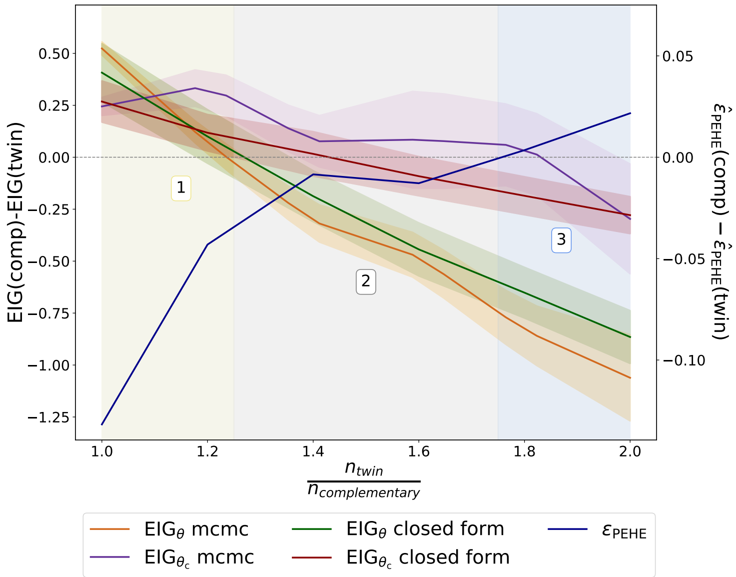


Figure 6.3: Difference in post host EIG and $\hat{\varepsilon}_{\text{PEHE}}$ for a linear CATE model trained on $\mathscr{D}_0 \cup \mathscr{D}_{comp}$ and $\mathscr{D}_0 \cup \mathscr{D}_{twin}$ for increasing $\frac{n_{twin}}{n_{comp}}$ and fixed $n_{host} = n_{comp} = 100$. PEHE is evaluated on hold-out data from the host distribution. The curves show mean $\pm$ 1 s.d. across 50 seeds. The three regions indicate different dataset preferences. In region 2, $\text{EIG}_{\theta|\mathscr{D}_0}$ incorrectly favors $\mathscr{D}_{twin}$, whereas $\text{EIG}_{\theta_c|\mathscr{D}_0}$ correctly selects $\mathscr{D}_{comp}$ for the causal task.

still the optimal in terms of CATE, but $\text{EIG}_{\theta|\mathscr{D}_0}$ preferences $\mathscr{D}_{twin}$ as it leads to a greater entropy reduction in the full set of parameters. This result shows that by focusing on the causal parameters alone, $\text{EIG}_{\theta_{c|\mathscr{D}_0}}$ is able to make the correct decision in selecting $\mathscr{D}_{comp}$. In the final region we see all lines have crossed the x axis, showing that all methods agree with the ground truth in choosing $\mathscr{D}_{twin}$. Finally, we note the MCMC estimates agree with the closed form counterparts, with increased variability due to sampling.

6.5.2 Ranking Experiment

Concept For our main experiment we validate our framework in a setting where the host must choose between many potential candidates, each with different distributions. To do so we begin with a standard causal inference benchmark dataset, $\mathscr{D}$, and form the host and candidates datasets using subsampling functions, $S_e(\mathbf{x},t)$ as detailed above where subsampling function is a logistic regression with random parameters. This ensures that each site has a different covariate distribution. We apply the three methods described in Section 6.3.3 to estimate both the two expected information gains for each candidate site, $\mathscr{D}_e$. Ultimately, like before, we compare the implied rankings with the ground truth ranking given by the PEHE. Full details are provided in section D.3 in the Appendix.

	$\rho(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial				
$\mathrm{EIG}_{\theta_c\|\mathscr{D}_0}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
$\mathrm{EIG}_{\theta\|\mathscr{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
PropScore Error	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
Causal GP				
$\mathrm{EIG}_{\tilde{\tau}(X_0)\|\mathscr{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
$\mathrm{EIG}_{\mathbf{f}\|\mathscr{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
Sample Size	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
Bayesian CF				
$\mathrm{EIG}_{\theta_c\|\mathscr{D}_0}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
$\mathrm{EIG}_{\theta\|\mathscr{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
PropScore Error	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$

Table 6.1: Ranking experiment for the IHDP dataset, measured by Spearman ρ and precision at k (p@k).

Baselines We provide a number of simple comparison methods as baselines for our task. Specifically, (a) ranking by sample size, (b) ranking by similarity of covariate distribution measured by a multivariate Gaussian fit to the host, and (c) ranking by dissimilarity of treatment allocation measured by the error of a propensity model fit on the host. We compare using Spearman rho(ρ) and precision at k (p@k).

Results Table 6.1 shows the results of our experiment on the IHDP dataset [169] where the host has to rank the best datasets out of 10 candidates. We report the average performance of the rankings across 20 repeated experiments, and according to Spearman ρ [186] and precision at k. We include the best baseline by Spearman ρ performance. Additional metrics and baseline performance can be found in section D.4. These results demonstrate that across all models, our EIG_{θ_c} based approach, focusing on causal parameters, outperforms the standard approach which seeks expected entropy reduction in all parameters. Further, our method works significantly better than all baselines.

Method	**MSE** $(\downarrow)$	$\rho(\uparrow)$
MPC Linear	$(\mathbf{3.80 \pm 0.04}) \times \mathbf{10^{-6}}$	$\mathbf{0.797 \pm 0.06}$
DP Linear	9.80 ± 0.30	0.06 ± 0.11

Table 6.2: Multi-Party Computation Results for EIG_{θ_c}.

6.5.3 Multi-Party Computation Experiments

Finally, we experimentally demonstrate the performance of our cryptographic protocol. We repeat a similar experiment as in Section 6.5.2 on the IHDP dataset, this time choosing between twenty datasets. Results are given using a linear model as this allows us to compare our MPC based approach against differential privacy [DP; 168, see section D.1.2 in the Appendix for definition]. We use $\varepsilon = 100$ and Laplace noising [51], following existing work on DP linear regression [187, 188]. In order to ensure a fair comparison, we noise the final statistics from the MPC computation with an appropriate amount of Laplace noise. This comes from the sensitivity of the EIG statistic which we derive in the Appendix under section D.1.3. Table 6.2 shows both the MSE induced by the privacy method and the Spearman ρ against the noise-free ranking. Results show that MPC vastly outperforms differential privacy, both in terms of MSE and ranking, with DP producing a near random ranking, despite the relaxed noising.

6.6 Related Work

Federated Learning Via Multi-Party Computation Our setting bares large with federated learning [189], a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Our goal differs from this as we focus not on learning a model across sites, but deciding which sites to pool. The application of multiparty computation within federated learning is a growing area [190, 191, 192], with much of the work focusing on how to learn predictive models in a secure cross site fashion. Most similar to our work is Muazu et al. [193], who develop a similar federated approach to data fusion focusing on healthcare, however they do not take a causal approach. To the best of our knowledge, there are no existing applications of multi-party computation within causal inference.

Federated/Private Causal Estimation There are, however, federated approaches to estimating causal effects. In this area, a majority of works partition the loss function into multiple components, with each component corresponding to a specific data source [194, 195, 196]. However, modelling complex, non-linear relationships remains challenging [197]. Many of these algorithms come without privacy guarantees, with the exception of Niu et al. [198], who add DP guarantees to various popular CATE estimation techniques. The latter approach however contrasts with ours as the use of sample splitting reduces data efficiency.

Causal Bayesian Active Learning Bayesian Active Learning by Disagreement [BALD; 172] is a framework for strategic training data acquisition, focusing on regions of high uncertainty. Our method based on maximising $\text{EIG}_{\theta|\mathscr{D}_0}$ (Equation 6.1) can be viewed as applying BALD after an initial update to an entire datasets rather than individual datapoints. Most similar to our work is CausalBALD [199], which applies an active learning approach to CATE estimation. However, the acquisition function cannot easily be extended to datasets without ignoring the correlation in information provided by different points. Most applications of active learning in causality focus on causal discovery and intervention selection [200, 201, 202].

6.7 Discussion

Limitations Due to the cost of computing high dimensional determinants and performing multiple rounds of conditional sampling, our method can be computationally costly in high dimensions. Moreover, the cost of multi-party computation will also increase as higher order approximations are required to retain accuracy. However, both of these remain negligible compared to the data engineering and expenses of data fusion [203, 204]. Finally, whilst our method offers a secure and principled way to prospectively quantify the value of dataset merges, the amount of information needed to justify such expenses may vary depending on application. It might, therefore, be beneficial to consider introducing a problem-specific threshold to determine when a proposed merger is worthwhile.

Conclusion We introduce an information-theoretic, cryptograpically secure framework for evaluating the utility of potential data merges for causal estimation. To the best of our knowledge, this is the first work addressing this relevant challenge. Through empirical evaluation, we demonstrate that our framework can reliably rank datasets according to their ability to improve CATE estimation. We show that entropy reduction in the CATE parameters alone gives an improvement when compared to a more standard approach to Bayesian dataset acquisition. Finally, we demonstrate that our cryptographic procedure can be applied in conjunction with DP, resulting in lower loss of accuracy, compared to applying DP alone.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Is merging worth it? Securely evaluating the information gain for causal dataset acquisition
Publication Status	□ Published X Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Fawkes, J.*, Ter-Minassian, L.*, Ivanova, D., Shalit, U. and Holmes, C., 2024. Is merging worth it? Securely evaluating the information gain for causal dataset acquisition. arXiv preprint arXiv:2409.07215. Accepted for publication at the AISTATS 2025 Conference

Student Confirmation

Student Name:	Lucile Ter-Minassian		
Contribution to the Paper	I am a joint first author of this paper with Jake Fawkes. Under the supervision of Prof. Holmes, I formulated the initial research question: how can we securely assess the value of a merge in causal estimation? Dr. Ivanova and I jointly developed the Bayesian Active Learning-inspired approach. Jake led the privacy component, including multi-party computation and the privacy experiment. Other experiments were designed and implemented collaboratively. While the paper was structured and written jointly, I took the lead in this aspect. The non-first authors provided guidance throughout and reviewed the manuscript.		
Signature	LTM	Date	18/02/2025

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments			
Signature	Chris Holmes	Date	25/02/2025

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 7

Conclusion

7.1 Summary

We have explored the fundamental challenges in approximating different types of truth in machine learning and causal inference. We have introduced methodological improvements to feature attribution methods that advance both local explainability and causal understanding, addressing critical gaps in the existing literature.

Our research began with Neighbourhood SHAP, which tackled the limitations of conventional Shapley analysis regarding locality. By leveraging local reference distributions rather than global ones, we improved on-manifold explainability and provided more meaningful feature attributions. This innovation directly addressed the tension between explanations that are faithful to the underlying data versus those that are faithful to the black-box model. Neighborhood Shapley values demonstrated increased robustness against adversarial classifiers and offered insights into local model behavior that global approaches could not capture.

We then bridged the gap between predictive modeling and causal understanding through the Path-Wise Shapley effects (PWSHAP) framework. By combining user-defined causal structures with Shapley values, we enabled practitioners to assess the targeted impact of binary variables through specific causal pathways. This approach allowed for clearer local bias and mediation analyses while remaining robust to adversarial attacks. PWSHAP reconciled both safety and reliability in feature attributions, providing a novel solution for explaining sensitive variables in complex models.

Our BICauseTree method advanced interpretable causal effect estimation by identifying clusters where natural experiments occur locally. Unlike post-hoc approaches to black-box models, our decision tree-based method offered inherent interpretability while improving covariate balance and reducing treatment allocation bias. The method's ability to detect subgroups with positivity violations and provide explicit covariate-based definitions of the target population enhanced both the transparency and generalizability of causal inferences.

Finally, we developed a cryptographically secure approach for quantifying the prospective value of data merges in causal estimation contexts. Our Expected Information Gain framework allowed data hosts to assess merge value without compromising sensitive information, addressing the critical challenge of reducing epistemic uncertainty and improving treatment overlap in privacy-constrained environments.

Throughout these contributions, we have consistently advanced the field toward more trustworthy AI by enhancing both explainability and causal understanding. Our methods collectively offer solutions that address the limitations of current approaches while providing practical tools for practitioners in high-stakes domains such as healthcare and policy-making.

7.2 Future Work

As the field of artificial intelligence evolves, particularly with the rise of foundation models, and in particular Large Language Models (LLMs), the focus of explainable AI has begun to shift toward global, mechanistic interpretability. This approach aims to dissect the underlying mechanisms of complex models rather than explaining individual predictions [205, 206]. While promising, this emerging direction faces significant challenges that warrant cautious consideration.

A key concern is the widespread reliance on the Linear Representation Hypothesis (LRH), which conjectures that neural networks represent high-level features or concepts as 'linear directions' within their activation space [207, 208]. The LRH is the fundamental assumption that drives many state-of-the-art methods in mechanistic interpretability—the field aiming to reverse engineer neural networks. For example, when trying to understand Large Language Models (LLMs), sparse autoencoders [209], linear probing [210], and steering vectors [211] all rely on the LRH. Christopher Olah [212] even remarks that for the field of mechanistic interpretability, *"if representations are not mathematically linear [...], it's back to the drawing board."* Therefore, as our best efforts to understand modern AI models relies upon the correctness of the LRH, it is critical for safe development of such systems to establish if it holds. Unfortunately, this fundamental assumption remains largely untestable in complex models, raising questions about the reliability of mechanistic interpretations derived from it.

Second, while mechanistic interpretability may offer insights into the general functioning of a model, it often falls short in justifying specific decisions for individuals [28]. This limitation is particularly problematic in contexts where individual fairness, accountability, and transparency are required. The methods we presented in this thesis, with their focus on local explanations and causal pathways, remain valuable complements to global interpretability approaches.

As models like GPT-4 [213] and Claude [214] are increasingly deployed in critical applications, we must exercise caution about using such sophisticated systems for important real-world decision-making when we cannot adequately explain their reasoning. The difficulty in interpreting LLMs stems not only from their scale but also from emergent capabilities that may not be predictable from training data alone [215].

The advances we have presented – from improved local explanations to causal understanding—provide stepping stones toward more interpretable and trustworthy AI systems. However, as model architectures evolve, so too must our explainability methods. Future work should focus on developing techniques that bridge the gap between local explainability and mechanistic interpretability, potentially by combining insights from both approaches to provide comprehensive explanations across different levels of abstraction.

In conclusion, while foundation models represent remarkable achievements in artificial intelligence, their deployment in high-stakes domains should be contingent on our ability to explain their decisions in ways that are meaningful to human stakeholders. Otherwise, we risk creating systems that, while powerful, remain fundamentally untrustworthy for critical applications.

Bibliography

[1] Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

[2] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

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

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

[5] Miguel A Hernán and James M Robins. *Causal Inference: What If.* Chapman & Hall/CRC, 2020.

[6] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[7] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[8] Francesca Dominici, Rachel C. Nethery, Fabrizia Mealli, and Jason D. Sacks. Causal inference and machine learning approaches for evaluation of the health impacts of large-scale air quality regulations. *arXiv preprint arXiv:1909.09611*, 2019.

[9] John Smith and Emily Doe. The impact of remote learning on academic outcomes during covid-19. *Journal of Educational Policy*, 45(2):123–145, 2021.

[10] Laura Johnson and Michael Lee. Optimizing hospital resource allocation during public health crises using causal modeling. *Health Policy and Planning*, 37(4): 567–579, 2022.

[11] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th*

ACM SIGKDD international conference on knowledge discovery & data mining, pages 3529–3530, 2020.

[12] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[13] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):118, 2020.

[14] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[15] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

[16] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[18] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[19] René F Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.

[20] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[21] Andrew Selbst and Julia Powles. “meaningful information” and the right to explanation. In *conference on fairness, accountability and transparency*, pages 48–48. PMLR, 2018.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[23] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[24] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[25] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[26] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3543–3554, 2018.

[27] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[28] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

[29] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *The Journal of Machine Learning Research*, 18(1):8753–8830, 2018.

[30] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[31] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[32] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[33] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.

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

[35] Maya L Petersen, Kristin E Porter, Susan Gruber, Yichuan Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

[36] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S84–S90, 2013.

[37] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1): 187–199, 2009.

[38] Angela Yaqian Zhu. *Causal Inference Methods for Addressing Positivity Violations and Bias in Observational and Cluster-Randomized Studies*. PhD thesis, University of Pennsylvania, 2021.

[39] Fan Li and Kari Lock Morgan. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[40] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[41] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113 (523):1228–1242, 2018.

[42] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

[43] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[44] Mark J. Van der Laan and Daniel Rubin. Targeted maximum likelihood estimation of a parameter of a marginal structural model. *The International Journal of Biostatistics*, 2(1), 2006.

[45] Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185 (1):65–73, 2017.

[46] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

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

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

[49] Daniel Westreich and Stephen R Cole. Invited commentary: positivity in practice. *American Journal of Epidemiology*, 171(6):674–677, 2010.

[50] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664, 2008.

[51] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[52] Deven McGraw, Sarah M Greene, Caroline S Miner, Karen L Staman, Mary Jane Welch, and Alan Rubel. Privacy and confidentiality in pragmatic clinical trials. *Clinical Trials*, 12(5):520–529, 2015.

[53] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Advances in Neural Information Processing Systems*, volume 33, pages 11637–11649, 2020.

[54] Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2020.

[55] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

[56] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[57] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[58] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[60] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635. ACM, 2021.

[61] Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

[62] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[63] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19. ACM, 2019.

[64] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[65] Krishna Gade, Sahin Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. Explainable ai in industry: Practical challenges and lessons learned. In *Companion Proceedings of the Web Conference 2020*, pages 303–304, 2020.

[66] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[67] Peyman Rasouli and Ingrid Chieh Yu. Meaningful data sampling for a faithful local explanation method. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 28–38. Springer, 2019.

[68] Tiago Botari, Frederik Hvilshøj, Rafael Izbicki, and Andre CPLF de Carvalho. Melime: Meaningful local explanation for machine learning models. *arXiv preprint arXiv:2009.05818*, 2020.

[69] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

[70] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR, 2020.

[71] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer, 2020.

[72] Giorgio Visani, Enrico Bagli, and Federico Chesani. Optilime: Optimized lime explanations for diagnostic computer algorithms. *arXiv preprint arXiv:2006.05714*, 2020.

[73] Adam White and Artur d'Avila Garcez. Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*, 2019.

[74] Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. Supervised neighborhoods for distributed nonparametric regression. In *Artificial Intelligence and Statistics*, pages 1450–1459. PMLR, 2016.

[75] Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, 2001.

[76] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[77] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[78] David Ruppert and Matthew P Wand. Multivariate locally weighted least squares regression. *The annals of statistics*, pages 1346–1370, 1994.

[79] Christopher Frye, Damien de Mijolla, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020.

[80] Cheng-Yu Hsieh, Chih-Kuan Yeh, Xuanqing Liu, Pradeep Ravikumar, Seungyeon Kim, Sanjiv Kumar, and Cho-Jui Hsieh. Evaluations and methods for explanation through robustness analysis. *arXiv preprint arXiv:2006.00442*, 2020.

[81] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[82] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

[83] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 2020.

[84] Tiago Botari, Rafael Izbicki, and Andre CPLF de Carvalho. Local interpretation methods to machine learning using the domain of the feature space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 241–252. Springer, 2019.

[85] Sean Saito, Eugene Chua, Nicholas Capel, and Rocco Hu. Improving lime robustness with smarter locality sampling. *arXiv preprint arXiv:2006.12302*, 2020.

[86] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and machine learning*, pages 159–175. Springer, 2018.

[87] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[88] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. On the (in) fidelity and sensitivity for explanations. *arXiv preprint arXiv:1901.09392*, 2019.

[89] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[90] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[91] Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 640–647, 2020.

[92] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Limeout: An ensemble approach to improve process fairness. *arXiv preprint arXiv:2006.10531*, 2020.

[93] Damien Garreau and Ulrike von Luxburg. Looking deeper into lime. *arXiv preprint arXiv:2008.11092*, 2020.

[94] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[95] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[96] Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580, 2004.

[97] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[98] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

[99] Peyman Rasouli and Ingrid Chieh Yu. Explan: Explaining black-box classifiers using adaptive neighborhood generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[100] Dilini Rajapaksha, Christoph Bergmeir, and Wray Buntine. Lormika: Local rule-based model interpretability with k-optimal associations. *Information Sciences*, 540: 221–241, 2020.

[101] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[102] Weifeng Liu, Jose C Principe, and Simon Haykin. *Kernel adaptive filtering: a comprehensive introduction*. John Wiley & Sons, 2011.

[103] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[104] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*, 2020.

[105] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021.

[106] Daniel Westreich and Sander Greenland. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

[107] Bertrand Iooss and Clémentine Prieur. Shapley effects for sensitivity analysis with dependent inputs: comparisons with sobol' indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9, 07 2017. doi: 10.1615/Int.J.UncertaintyQuantification.2019028372.

[108] Guido W Imbens and Donald B Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

[109] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[110] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2021.

[111] Raghav Singal, George Michailidis, and Hoiyi Ng. Flow-based attribution in graphical models: A recursive Shapley approach. In *International Conference on Machine Learning*, pages 9733–9743. PMLR, 2021.

[112] Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 33:10999–11009, 2020.

[113] Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.

[114] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[115] Tong Wang and Cynthia Rudin. Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*, 2017.

[116] Chris C Holmes and James A Watson. Machine learning for randomised controlled trials: identifying treatment effect heterogeneity with strict control of type I error. *bioRxiv*, page 330795, 2018.

[117] Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509, 2020.

[118] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

[119] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[120] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is “nearest neighbor” meaningful? In *Database Theory—ICDT’99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7*, pages 217–235. Springer, 1999.

[121] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

[122] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[123] Ehud Karavani, Peter Bak, and Yishai Shimoni. A discriminative approach for finding and characterizing positivity violations using decision trees. *arXiv preprint arXiv:1907.08127*, 2019.

[124] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

[125] Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

[126] Guy Wolf, Gil Shabat, and Hanan Shteingart. Positivity validation detection and explainability via zero fraction multi-hypothesis testing and asymmetrically pruned decision trees. *arXiv preprint arXiv:2111.04033*, 2021.

[127] Samuel Ackerman, Eitan Farchi, Orna Raz, Marcel Zalmanovici, and Parijat Dube. Detection of data drift and outliers affecting machine learning model performance over time. *arXiv preprint arXiv:2012.09258*, 2020.

[128] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[129] Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

[130] Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203, 1973.

[131] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[132] Hervé Abdi. Holm's sequential bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8, 2010.

[133] Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. 2007.

[134] John R Quinlan et al. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence*, volume 92, pages 343–348. World Scientific, 1992.

[135] Donald B Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, pages 293–298, 1980.

[136] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[137] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 20(1):1–24, 2012.

[138] Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

[139] P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.

[140] Rom Gutman, Ehud Karavani, and Yishai Shimoni. Propensity score models are better when post-calibrated. *arXiv preprint arXiv:2211.01221*, 2022.

[141] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[142] Department of Health and Human Services, Food and Drug Administration. Using artificial intelligence and machine learning in the development of drug and biological products; availability. `https://www.fda.gov/media/167973/download?attachment`, 2023. [Docket No. FDA–2023–N–0743].

[143] Aniek Erasmus, Tyler DP Brunet, and Eran Fisher. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, 47(5):329–335, 2021.

[144] John Kellett, Lucien Wasingya-Kasereka, Mikkel Brabrand, and Martin Opio. Two simple replacements for the triage early warning score to facilitate the south african triage scale in low resource settings. *African Journal of Emergency Medicine*, 11 (1):45–51, 2021.

[145] Habiba Muhammad Sani, Ci Lei, and Daniel Neagu. Computational complexity analysis of decision tree algorithms. In *Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence, AI 2018, Cambridge, UK, December 11–13, 2018, Proceedings 38*, pages 191–197. Springer, 2018.

[146] Merle Behr, M Azim Ansari, Axel Munk, and Chris Holmes. Testing for dependence on tree structures. *Proceedings of the National Academy of Sciences*, 117(18):9787–9792, 2020.

[147] Michael Patrick Allen. Testing hypotheses in nested regression models. *Understanding regression analysis*, pages 113–117, 1997.

[148] Alan Malek, Sumeet Katariya, Yinlam Chow, and Mohammad Ghavamzadeh. Sequential multiple hypothesis testing with type i error control. In *Artificial Intelligence and Statistics*, pages 1468–1476. PMLR, 2017.

[149] Avi Feller and Andrew Gelman. Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource*, pages 1–16, 2015.

[150] Federico Castanedo et al. A review of data fusion techniques. *The scientific world journal*, 2013, 2013.

[151] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[152] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.

[153] Jodyn Platt and Sharon Kardia. Public trust in health information sharing: implications for biobanking and electronic health record systems. *Journal of personalized medicine*, 5(1):3–21, 2015.

[154] Michelle M Mello, Jeffrey K Francer, Marc Wilenzick, Patricia Teden, Barbara E Bierer, and Mark Barnes. Preparing for responsible sharing of clinical trial data, 2013.

[155] European Commission. Data protection rules for the public sector. Retrieved from `https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-public-sector_en`, 2018.

[156] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. Big healthcare data: preserving security and privacy. *Journal of big data*, 5(1):1–18, 2018.

[157] Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

[158] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[159] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

[160] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[161] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[162] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

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

[164] Andrew C Yao. Protocols for secure computations. In *23rd annual symposium on foundations of computer science (sfcs 1982)*, pages 160–164. IEEE, 1982.

[165] David Evans, Vladimir Kolesnikov, Mike Rosulek, et al. A pragmatic introduction to secure multi-party computation. *Foundations and Trends® in Privacy and Security*, 2(2-3):70–246, 2018.

[166] Brian Knott, Shobha Venkataraman, Awni Hannun, Shubho Sengupta, Mark Ibrahim, and Laurens van der Maaten. Crypten: Secure multi-party computation meets machine learning. *Advances in Neural Information Processing Systems*, 34:4961–4973, 2021.

[167] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

[168] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[169] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[170] Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

[171] Judea Pearl. Causal inference in statistics: An overview. 2009.

[172] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[173] Paola Sebastiani and Henry P Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.

[174] Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *International Conference on Machine Learning*, pages 4267–4276. PMLR, 2018.

[175] Adam Evan Foster. *Variational, Monte Carlo and policy-based approaches to Bayesian experimental design*. PhD thesis, University of Oxford, 2021.

[176] Alan Goldstein, Alan E. Gelfand, and D. David Shafer. Bayesian nonparametric modeling for causal inference. *Journal of the American Statistical Association*, 106 (494):263–276, 2011.

[177] V. V. Fedorov. Theory of optimal experiments. In *Academic Press*, 1972.

[178] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, 6(1):266–298, 2012.

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

[180] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3): 195–266, 2012.

[181] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[182] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[183] Donald Beaver. Efficient multiparty protocols using circuit randomization. In *Advances in Cryptology—CRYPTO'91: Proceedings 11*, pages 420–432. Springer, 1992.

[184] Alston Scott Householder. The numerical treatment of a single nonlinear equation. *(No Title)*, 1970.

[185] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[186] Charles Spearman. The proof and measurement of association between two things. 1961.

[187] Garrett Bernstein and Daniel R Sheldon. Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

[188] Jordan Awan and Aleksandra Slavković. Structure and sensitivity in differential privacy: Comparing k-norm mechanisms. *Journal of the American Statistical Association*, 116(534):935–954, 2021.

[189] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

[190] Yong Li, Yipeng Zhou, Alireza Jolfaei, Dongjin Yu, Gaochao Xu, and Xi Zheng. Privacy-preserving federated learning framework based on chained secure multiparty computing. *IEEE Internet of Things Journal*, 8(8):6178–6186, 2020.

[191] Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, and Tucker Hybinette Balch. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*, volume 21. MIT Press Cambridge, MA, USA, 2019.

[192] Renuga Kanagavelu, Zengxiang Li, Juniarto Samsudin, Yechao Yang, Feng Yang, Rick Siow Mong Goh, Mervyn Cheah, Praewpiraya Wiwatphonthana, Khajonpong Akkarajitsakul, and Shangguang Wang. Two-phase multi-party computation enabled privacy-preserving federated learning. In *2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID)*, pages 410–419. IEEE, 2020.

[193] Tasiu Muazu, Yingchi Mao, Abdullahi Uwaisu Muhammad, Muhammad Ibrahim, Umar Muhammad Mustapha Kumshe, and Omaji Samuel. A federated learning system with data fusion for healthcare using multi-party computation and additive secret sharing. *Computer Communications*, 216:168–182, 2024.

[194] Thanh Vinh Vo, Arnab Bhattacharyya, Young Lee, and Tze-Yun Leong. An adaptive kernel approach to federated learning of heterogeneous causal effects. *Advances in Neural Information Processing Systems*, 35:24459–24473, 2022.

[195] Yuxuan Liu, Haozhao Wang, Shuang Wang, Zhiming He, Wenchao Xu, Jialiang Zhu, and Fan Yang. Disentangle estimation of causal effects from cross-silo data. *arXiv preprint arXiv:2401.02154*, 2024.

[196] Thanh Vinh Vo, Tze-Yun Leong, et al. Federated learning of causal effects from incomplete observational data. *arXiv preprint arXiv:2308.13047*, 2023.

[197] Alejandro Almodóvar, Juan Parras, and Santiago Zazo. Federated learning for causal inference using deep generative disentangled models. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023.

[198] Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially private estimation of heterogeneous causal effects. In *Conference on Causal Learning and Reasoning*, pages 618–633. PMLR, 2022.

[199] Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478, 2021.

[200] Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peharz, and Julius Von Kügelgen. Active bayesian causal inference. *Advances in Neural Information Processing Systems*, 35:16261–16275, 2022.

[201] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

[202] Yashas Annadani, Panagiotis Tigas, Desi R Ivanova, Andrew Jesson, Yarin Gal, Adam Foster, and Stefan Bauer. Differentiable multi-target causal bayesian experimental design. *arXiv preprint arXiv:2302.10607*, 2023.

[203] Michael L Brodie. Data integration at scale: From relational data integration to information ecosystems. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 2–3. IEEE, 2010.

[204] Anirudh Kadadi, Rajeev Agrawal, Christopher Nyamful, and Rahman Atiq. Challenges of data integration and interoperability in big data. In *2014 IEEE international conference on big data (big data)*, pages 38–40. IEEE, 2014.

[205] Neel Nanda, Luca Lieberum, Mehran Haghifam, Johannes Musser, Dishant Magar, Jiri Berrevoets, John Jumper, Jacob Steinhardt, Noah Goodman, Tania Lombrozo, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *arXiv preprint arXiv:2402.14705*, 2023.

[206] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[207] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[208] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[209] Trenton Bricken, Siva Reddi, Alex Efrat, Scott Gray, Thomas L Griffiths, Ethan Odell, and Jacob Steinhardt. Monosemanticity: Learning semantically disentangled features with generative models. *arXiv preprint arXiv:2310.17230*, 2023.

[210] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[211] Alex Turner, Domenic Gurnee, Linden Agarwal, Jacob Chanenson, Hailey Bernstein, Laura Seethor, Jeffrey Lin, Jordan Ash, Morris Perelstein, Hans Bakker, et al. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[212] Chris Olah. Getting ai alignment right might be hard, don't expect the first 10 plans to work. EleutherAI Alignment Reading Group, 2024.

[213] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[214] Anthropic. Claude: Technical report. Technical report, Anthropic, 2023. URL `https://www.anthropic.com/claude-technical-report`.

[215] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[216] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. Model agnostic supervised local explanations. *arXiv preprint arXiv:1807.02910*, 2018.

[217] Art B Owen. Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

[218] Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation via linear regression. *arXiv preprint arXiv:2012.01536*, 2020.

[219] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[220] Tyler J VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468, 2009.

[221] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1287–1297, 2021.

[222] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

[223] Numair Sani, Daniel Malinsky, and Ilya Shpitser. Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*, 2020.

[224] Maya L Petersen and Mark J van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418, 2014.

[225] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[226] Yang Song and George YH Chi. A method for testing a prespecified subgroup in clinical trials. *Statistics in medicine*, 26(19):3535–3549, 2007.

[227] Anton Nilsson, Carl Bonander, Ulf Strömberg, and Jonas Björk. Assessing heterogeneous effects and their determinants via estimation of potential outcomes. *European journal of epidemiology*, 34(9):823–835, 2019.

[228] Han Wu, Sarah Tan, Weiwei Li, Mia Garrard, Adam Obeng, Drew Dimmery, Shaun Singh, Hanson Wang, Daniel Jiang, and Eytan Bakshy. Distilling heterogeneity: From explanations of heterogeneous treatment effect models to interpretable policies. *arXiv preprint arXiv:2111.03267*, 2021.

[229] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

[230] Guido W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

[231] Paul R Rosenbaum. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*, 2005.

[232] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[233] Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

[234] Tyler J VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

[235] Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975.

[236] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

[237] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[238] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[239] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[240] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

[241] Peter C Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics*, 10(2):150–161, 2011.

[242] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political analysis*, 27(4):435–454, 2019.

[243] Shaun R Seaman and Stijn Vansteelandt. Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184, 2018.

[244] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[245] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Ann. Statist*, 47(2):1148–1178, 2019.

[246] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML, 2019. Version 0.x.

[247] Richard K Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

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

[249] Roman V Yampolskiy. *Artificial intelligence safety and security*. CRC Press, 2018.

[250] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3):231–237, 2019.

[251] Jingyu He, Saar Yalov, and P Richard Hahn. Xbart: Accelerated bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138. PMLR, 2019.

[252] Nikolay Krantsevich, Jingyu He, and P Richard Hahn. Stochastic tree ensembles for estimating heterogeneous effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6120–6131. PMLR, 2023.

[253] Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

[254] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

[255] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[256] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[257] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.

[258] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE conference on computer communications*, pages 1698–1707. IEEE, 2020.

[259] Shinji Tarumi, Mayumi Suzuki, Hanae Yoshida, Shoko Miyauchi, and Ryo Kurazume. Personalized federated learning for institutional prediction model using electronic health records: A covariate adjustment approach. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE, 2023.

Appendix A

Supplementary Materials: On Locality of Local Explanation Models

A.1 Local Linear Approximations vs Additive Feature Attribution Methods: further details

To illustrate the difference between local linear approximations and additive explanation models, consider explaining a black box $f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$ given i.i.d. realisations $\{x_i\}_{i=1}^N$ of $X \sim \text{Normal}(0,1)$. Results are depicted in Figure A1. As we see the attribution of Feature-1 computed by SHAP and LIME is constant no matter how far we are from the decision boundary. This result is expected since for fixed Feature-2 the change in model outcome is constant for removing Feature-1 whenever Feature-1 is positive (or negative respectively). [216] define *locality* as the aim to understand a prediction by asking 'if the input is changed slightly, how does the model's prediction change?'. As this example shows, LIME and SHAP capture global patterns of the black box model, despite being referred to as 'local' explanation models. This observation motivates the use of Neighbourhood SHAP. A linear approximation model such as MeLIME fits a linear approximation around x. Since at small negative values for Feature-1, observations with a positive value for Feature-1 are also included in the neighbourhood, the linear approximation is influenced by the higher positive effect of the black box and thus attributes the feature positively.

Note that LIME is a general framework that has been defined as additive feature attribution method (also see Supplement A.5). However, its implementation allows to also train a linear approximation by sampling data from the reference distribution, weighting it with its distance to the test instance x, and training a linear model such as Ridge on the weighted data. Some extensions such as MeLIME [68] or MAPLE [67] extend LIME by proposing more meaningful neighbourhood sampling schemes. However they only consider a linear approximation setting. It has been noted that LIME underperforms compared to linear

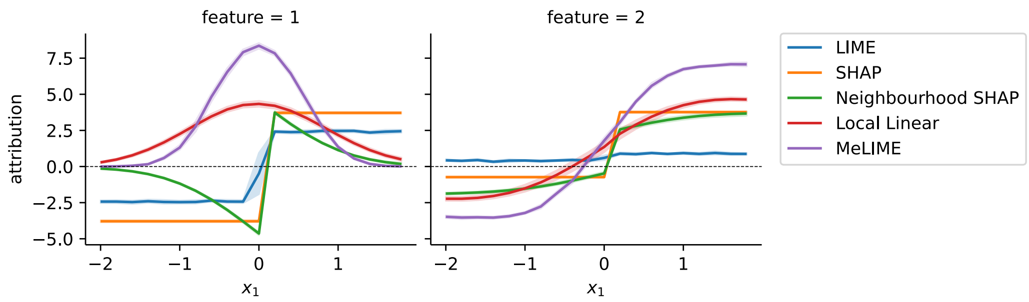


Figure A1: Attributions at $x = (x_1, 2)$ with x_1 varying for a reference distribution of $X \sim \text{Normal}(0,1)$ and black box $f(x) = \mathbb{I}(x_1 > 0)2x_2^2 - \mathbb{I}(x_1 \leq 0)x_2^2$ averaged over 10 runs displayed with 95% confidence intervals (see next section for details). While (Tabular) LIME and SHAP assign the same absolute attribution to Feature-1 no matter how large x_1 is, our neighbourhood approach takes its distance to the decision boundary into consideration. A local linear approximation to the black box trained with an euclidean distance based exponential kernel weighted Ridge Regressor, i.e. on the same neighbourhood as Neighbourhood SHAP, gives misleading attributions to Feature-1 for $-0.4 < x_1 < 0$.

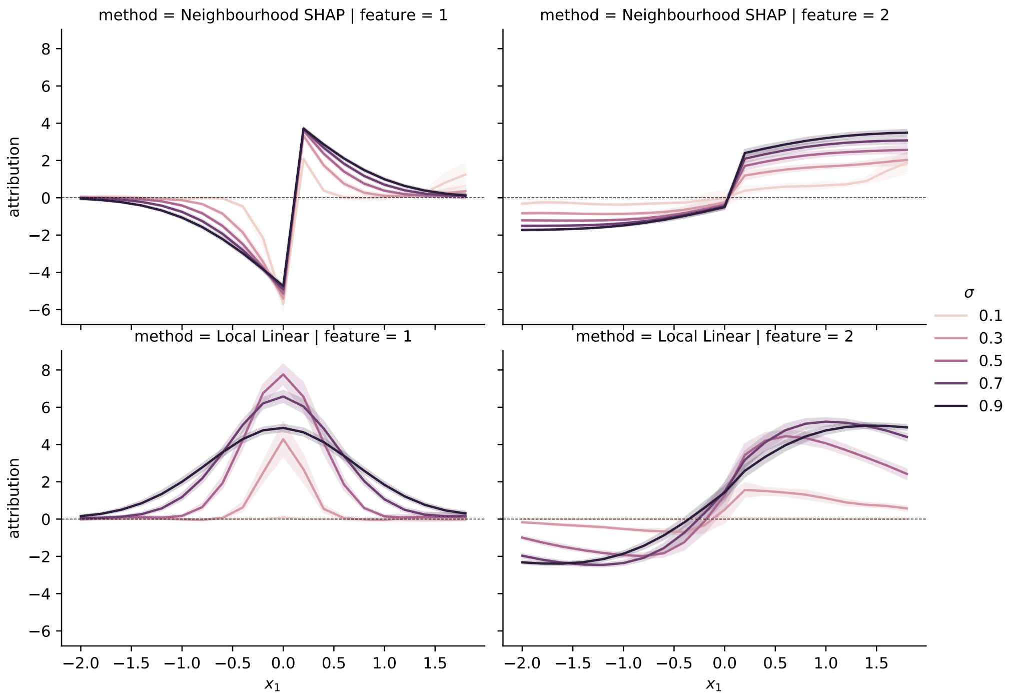


Figure A2: Feature attributions for the earlier example for different bandwidths (differentiated by the hue) for Neighbourhood SHAP (first row) and a local linear approximation (second row). As we see, the local linear model has a higher variance in general. We also note that the local linear attribution of Feature-2 goes to 0 for larger values of x_1 which is not desired. This results from the fact that neighbourhoods around x are sparser the larger x gets. For $\sigma = 0.1$, we note that Neighbourhood SHAP also suffers under high variance in the Feature-2 attribution for extreme values.

approximation methods [68, 216] when performance is measured by prediction accuracy within a small neighbourhood around x. This is not surprising as the Tabular version of LIME as additive feature attribution method (which the papers compare to) is not defined by taking local neighbourhoods into consideration.

In contrast to local linear approximations, our proposal, Neighbourhood SHAP, increases the 'locality' of SHAP by sampling the removed features from a neighbourhood around x, instead of from the (global) statistical population. There are several reasons why a user might be more interested in an additive feature attribution (referred to as method [B]) that builds upon neighbourhood distributions than a local linear approximation (referred to as method [A]):

- Approach [A] requires the subjective choice of the parametric local model $g(x)$. As has been pointed out elsewhere [28], if $g(x)$ is a good approximation to $f(x)$ for all $x \sim X$ then the user should adopt $g(x)$ as their preferred model (as it is explainable). If $g(x)$ is not a faithful local representation of $f(x)$ then it is challenging to trust its interpretability.

- Approach [A] translates the local behaviour of $f(x)$ through $g(x)$ and hence any interpretation and statement of explanations must be contextualised in light of this, as the translator's version of $f(x)$. Instead approach [B], and in particular Neighbourhood SHAP, has an interpretation in terms of the expected change in $f(x)$ that does not require a surrogate model.

- Approach [B] is model free, i.e. it does not make any parametric assumptions. The method attributes changes in the local expectations of $f(x)$ to features of x through a sum-of-squares (variance) decomposition [217].

Taken together we believe approach [B] has strong merit as a local explainability measure.

A.2 Axioms of Shapley Values

Shapley values have been shown to satisfy the following axioms.

- According to the *Dummy* axiom, a feature j receives a zero attribution if it has no possible contribution, i.e. $v(S \cup j) = v(S)$ for all $S \subseteq \{1,...,m\}$.

- According to the *Symmetry* axiom, two features that always have the same contribution receive equal attribution, i.e. $v(S \cup i) = v(S \cup j)$ for all S not containing i or j then $\phi_i(v) = \phi_j(v)$.

- According to the *Efficiency* axiom, the attributions of all features sum to the total value of all features. Formally, $\sum_j \phi_j(v) = v(\{1,..,m\})$.
- According to the *Linearity* axiom, for any value function v that is a linear combination of two other value functions u and w (i.e. $v(S) = \alpha u(S) + \beta w(S)$), the Shapley values of v are equal to the corresponding linear combination of the Shapley values of u and w (i.e. $\phi_i(v) = \alpha\phi_i(u) + \beta\phi_i(w)$).

Since all these axioms have been defined conditionally on the value function v, they hold if the value function is changed. We have expressed both Neighbourhood SHAP and Smoothed SHAP as a change in the value function of standard Shapley values. As such they also adhere to the axioms.

A.3 Computational Burden

In this section, we want to illustrate the computational complexity of Neighbourhood SHAP. Please refer to Figures A3 for plots that describe the additional computational complexity of computing the proposed SHAP values.

Computation using the Shapley formula. One way of computing Shapley values is to use a mean estimator. For marginal reference distributions, it holds in particular that

$$\phi(j,x) = \mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x_S)} \left[\mathop{\mathbb{E}}_{S}[f(x_{S\cup j}, X^*_{\overline{S}\setminus j}) - f(x_S, X^*_{\overline{S}})] \right] = \mathop{\mathbb{E}}_{x^* \sim r(X^*_{\overline{S}} \mid x_S)} [\phi_{x^*}(j)] .$$

Then,

$$\phi_{x^*}(j) = \mathop{\mathbb{E}}_{S}[f(x_{S\cup j}, x^*_{\overline{S}\setminus j}) - f(x_S, x^*_{\overline{S}})] \tag{A.1}$$

is referred to as single-reference Shapley value [71], which can be characterised as the expected change in model outcome when feature j of observation x^* is set equal to x_j after a random subset of features x^*_S has already been set equal to x_S. When the reference distribution is marginal, we can write the local Shapley value as a weighted average of the single-reference Shapley values

$$\hat{\phi}^{local}(j) = \frac{1}{\sum_{i=1}^{L} d(x^*_{i,\overline{S}} \mid x_{\overline{S}})} \sum_{i=1}^{L} d(x^*_{i,\overline{S}} \mid x_{\overline{S}}) \phi_{x^*_i}(j)$$

where $\{x^*_i\}_{i=1}^{L}$ are samples from $r(x^* \mid x)$. Computing Neighbourhood SHAP is then just a linear transformation of the single-reference Shapley values. As such the complexity for computing Neighbourhood SHAP for any additional σ_N for all features is $\mathcal{O}(L \cdot m)$: after the computation of L weights, these have to multiplied with the single-reference Shapley values and the weighted values are added up.

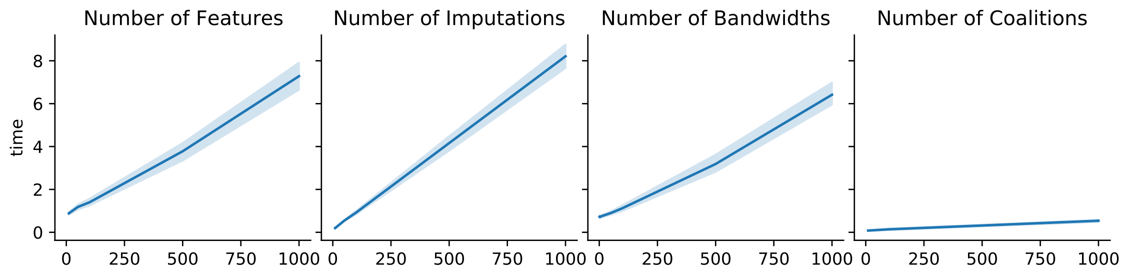


Figure A3: Plots of the computational time (seconds in clock time) to compute Neighbourhood SHAP w.r.t the number of imputations L (default: 100), the number of features m (default: 11), the number of coalitions C (default: 2^{11}), and the number of bandwidths (default: 50) averaged over 10 runs displayed with 95% confidence intervals, run on a 2.4 GHz 6-Core Intel Core i5-9300H CPU, using the SHAP package [57].

Computation using KernelSHAP. The KernelSHAP optimisation of

$$\mathbb{E}_S[(\mathbb{E}_{r(X^*_{\bar{S}} \mid x_S)}[f(x_S, X^*_{\bar{S}})] - g(S))^2] \approx \sum_{l=1}^{L}\sum_{i=1}^{C} w_i(f(x_{S_i}, x^*_{l,\bar{S}_i}) - g(S_i))^2 \tag{A.2}$$

can be simplified to

$$\mathbb{E}_S[(\mathbb{E}_{r(X^*_{\bar{S}} \mid x_S)}[f(x_S, X^*_{\bar{S}})] - g(S))^2] \approx \sum_{i=1}^{C} w_i(\frac{1}{L}\sum_{l=1}^{L} f(x_{S_i}, x^*_{l,\bar{S}_i}) - g(S_i))^2. \tag{A.3}$$

As such we only add weighting to the optimisation

$$\mathbb{E}_S[(\mathbb{E}_{r(X^*_{\bar{S}} \mid x_S)}[f(x_S, X^*_{\bar{S}})] - g(S))^2] \approx \sum_{i=1}^{C} w_i(\sum_{l=1}^{L} \frac{d(x^*_l \mid x)}{\sum_{k=1}^{L} d(x^*_k \mid x)} f(x_{S_i}, x^*_{l,\bar{S}_i}) - g(S_i))^2. \tag{A.4}$$

Computing the weights can be done in $\mathcal{O}(L \cdot m)$, while the cost for aggregation of the model outcome over coalitions is an additional $\mathcal{O}(L \cdot C)$. Finally, the optimisation using matrix multiplication is of complexity $\mathcal{O}(C \cdot m)$. All in all, the computational cost is thus $\mathcal{O}(\max(L, m) \cdot C)$.

Time Complexity of Smoothed SHAP. Once N Shapley values have been computed, Smoothed SHAP can be computed in $\mathcal{O}(N \cdot m)$.

A.4 Anti-Neighbourhood SHAP

Among others, Shapley values satisfy the linearity axiom [57] which says that two attributions $\phi_v(j), \phi_w(j)$ with value functions v, w add up to $\phi_{v+w}(j)$. As a result, we can characterise the difference between the marginal Shapley values $\phi_{\mathbb{E}_{r(X^* \mid x_S)}\left[f(x_S, X^*_{\bar{S}})\right]}$ and the

Neighbourhood Shapley values $\phi_{\mathbb{E}_{n(X^* \mid x)}\left[f(x_S, X^*_{\bar{S}})\right]}$ as Anti-Neighbourhood Shapley values $\phi^{anti} = \phi_{\mathbb{E}_{r(X^* \mid x_S)}\left[f(x_S, X^*_{\bar{S}})\right] - \mathbb{E}_{n(X^* \mid x)}\left[f(x_S, X^*_{\bar{S}})\right]}$

$$\mathop{\mathbb{E}}_{r(x^* \mid x)}\left[f(x_S, x^*_{\bar{S}})\right] - \mathop{\mathbb{E}}_{n(x^* \mid x)}\left[f(x_S, x^*_{\bar{S}})\right] \approx \frac{1}{L}\sum_{i=1}^{L} f(x_S, x^*_{i,\bar{S}})\left(1 - \frac{d(x^*_{i,\bar{S}} \mid x_{\bar{S}})}{\frac{1}{L}\sum_{l=1}^{L} d(x^*_{l,\bar{S}} \mid x_{\bar{S}})}\right).$$

While Neighbourhood SHAP computes the expected change in model outcome when the features of a random observation in the neighbourhood of x are set equal to x, Anti-neighbourhood SHAP weights the change in model outcome higher for observations that are farther away from x. This might be of interest, if the user is worried about loss in information from only using Neighbourhood SHAP. Consider the black box from Supplement A.1. Even though Feature-1 has no local effect on the black box for large values of Feature-1, it affects the value of the test instance globally: this is reflected by Anti-Neighbourhood SHAP.

A.5 Explaining LIME

LIME [22] is defined as an optimisation problem in the binary coalition space

$$\varepsilon(x) = \text{argmin} \sum_{z, z' \in \mathscr{Z}} d(z|x)(f(z) - g(z'))^2 + \Omega(g) \tag{A.5}$$

where $\Omega(g)$ is a penalty on the complexity of g, i.e. an $l1$ loss, z are samples drawn from the feature space, and z' are the corresponding binary representations. While the computation of LIME differs depending on the data type (i.e. image, text or tabular data), we will focus on tabular data here, as its corresponding implementation is used for comparisons on simulated data sets [68] or to determine predictive accuracy [67].
In order to compute the attribution at a local observation x, Tabular LIME follows the following steps:

1. Given a background data set, learn the k quantiles $q_1, ..., q_k$ of the data distribution with their summary statistics (mean and standard deviation).

2. Compute each dimension j of each imputation i, $z_{i,j}$, with $i \in \{1, ..., L\}, j \in \{1, ..., m\}$ as follows

   a) Sample a quantile $b_{i,j}$ for each feature j by uniformly sampling from $\{q_1, ..., q_k\}$.

   b) Sample an observation $z_{i,j}$ from a truncated Gaussian fitted with the summary statistics of quantile $b_{i,j}$ for each feature j.

c) Turn the observation z_i into its discretised mapping z_i' by setting $z_{i,j}$ equal to 1 if feature j of the test instance, i.e. x_j, falls in quantile $b_{i,j}$, and 0 otherwise.

3. Solve the optimisation problem in Equation A.5 with $d(z|x)$ replaced by $d(z'|x')$ where $x' = 1_m$ is a vector of all ones as x_j always falls in the same bin as x_j.

For a more in depth depiction of the algorithm, we refer the reader to [93]. Note that any observation u where each dimension falls into the same bin as the corresponding dimension from x, i.e. $q(u_j) = q(x_j)$ where q returns the quantile of its argument, gets the same attribution as x. As such, LIME is an aggregated attribution measure. However only observations that fall into the same bin in each dimension get the same attribution. As such it uses naive histogram weights, i.e. $I(q(z_j) = q(x_j)$ for all $j)$ To analyse the smoothness of the LIME attributions, let the lower bound of q be denoted by q_l while the upper bound is denoted by q_u. Now consider two observations x and u with $x_j = q_{1,l} + \frac{1}{2}\sqrt{\varepsilon/m}$ and $u_j = q_l(x_j) - \frac{1}{2}\sqrt{\varepsilon/m}$ for each feature j for an arbitrary ε. It then follows that $\|x-u\| \leq \varepsilon$. Since we cannot make any statement about the difference in attributions of any two such observations without additional assumptions, Lipschitz continuity does not have to hold necessarily. Note that a difference to Smoothed SHAP is that the smoothing is applied to each dimension independently.
Some other interesting properties of this algorithm include:

- The reference values z are sampled from a global reference distribution which consists of a mixture of non-overlapping truncated Gaussian distributions where each mixture component has the same probability.
- Each dimension of z is sampled independently.
- The probability of a coalition z', or as earlier defined by S, is:
  $P(S) \propto \exp(-(|S|-m)^2/\sigma^2)/k^{|S|}$ where k is the number of quantiles.

Note that often LIME is understood as sampling local data by weighting it with a distribution that depends on the distance in the data space [68, 85]. However, interestingly enough this is not how Tabular LIME is implemented by [22].
Instead, observations with a close coalition representation z' are weighted higher. This does not ensure that close observations receive a higher importance. Consider a three dimensional feature space and a test instance of $x = (1,1,1)$, the model outcome $f(1,1,10000)$ for coalition $S = \{1,2\}$ will be weighted higher in the optimisation than the model outcome $f(1,2,2)$ for coalition $S = \{1\}$. This enforced locality in the coalition space of LIME will lead to lower attributions of a feature if its effect is reduced by the presence of another feature. Consider for instance a simple rule based model such as $f(x) = \mathbb{I}(x_1 > 1 \text{ or } x_2 > 1)x_3$

Feature	1	2	3
LIME	0.028	0.018	0.624
SHAP	0.090	0.090	0.819
(Ours) Neighbourhood SHAP ($\sigma = 0.5$)	0.063	0.074	0.079

Table A1: Attributions at $x = (1.001, 1.001, 1.001)$ for a reference distribution of $X \sim \text{Uniform}(-2, 2)$ and black box $f(x) = \mathbb{I}(x_1 > 1 \text{ or } x_2 > 1)x_3$ computed over 10 runs.

at $x = (1.001, 1.001, 1.001)$. Perturbing, *either* Feature-1 *or* Feature-2 does not have any effect on the model outcome. For a reference distribution of $X \sim \text{Uniform}(-2, 2)$, the attributions of x_1 and x_2 are therefore almost zero (Table A1). In contrast, Shapley values find a relatively higher attribution for Features-1 and 2, but still a considerably higher attribution for Feature-3. In a small neighbourhood around x, this attribution can be misleading.

A.6 Nadaraya-Watson Estimator

We show that the Nadaraya-Watson estimator can be interpreted as an importance sampling estimator. In kernel regression the aim is to model the non-linear relationship between a dependent variable Y and an independent variable Z, by approximating the conditional expectation $\mathbb{E}[Y \mid Z]$. This is traditionally argued for by an approximation of the conditional density $p(y|z)$ by estimating both $p(z, y)$ and $p(z)$ with kernel regressors, i.e.

$$\mathbb{E}[Y|Z = z] = \int y \cdot p(y|z)dy = \int y \frac{p(z,y)}{p(z)} dy \text{ with estimators } \hat{p}(x) = \frac{1}{L}\sum_{i=1}^{L} d(z_i|z),$$

$$\hat{p}(z,y) = \frac{1}{L}\sum_{i=1}^{L} d(z_i|z)d(y_i|y) \text{ and thus } \hat{\mathbb{E}}[Y|Z = z] = \frac{\sum_{i=1}^{L} d(z_i|z)y_i}{\sum_{j=1}^{L} d(z_j|z)}.$$

A.7 Lipschitz Continuity of Smoothed SHAP

For a data set $\{x_i\}_{i=1}^{N}$ and bounded black box f, we show that Smoothed SHAP increases Lipschitz continuity. To ensure this, for any two values $z, y \in \mathbb{R}^m$ with $k \in \{1, \ldots, m\}$ it has to hold that

$$\left\| \hat{\phi}^{smoothed}(k,z) - \hat{\phi}^{smoothed}(k,y) \right\|^2 \leq L^2 \left\| z - y \right\|^2$$

for a constant L. Note that not any bounded function is Lipschitz continuous, and that this is thus a non-trivial proof.

For simplicity we drop the index k in the following. Let $\delta^2 = \|z - y\|^2$. Because of the triangle inequality it follows that $\|x_i - z\|^2 \leq \|x_i - y\|^2 + \delta^2$. Building upon this, we derive that

$$
\begin{aligned}
&\left\|\hat{\phi}^{smoothed}(z) - \hat{\phi}^{smoothed}(y)\right\| \\
&= \left\| \frac{1}{\sum_{j=1}^{N} d(x_j \mid z)} \sum_{i=1}^{N} \hat{\phi}(x_i) d(x_i \mid z) - \frac{1}{\sum_{j=1}^{N} d(x_j \mid y)} \sum_{i=1}^{N} \hat{\phi}(x_i) d(x_i \mid y) \right\| \\
&= \left\| \sum_{i=1}^{N} \hat{\phi}(x_i) \left(\frac{d(x_i \mid z)}{\sum_{j=1}^{N} d(x_j \mid z)} - \frac{d(x_i \mid y) \sum_{j=1}^{N} d(x_j \mid z)}{\sum_{j=1}^{N} d(x_j \mid y) \sum_{j=1}^{N} d(x_j \mid z)} \right) \right\| \\
&= \left\| \sum_{i=1}^{N} \hat{\phi}(x_i) \left(\frac{\exp(-\|x_i - z\|^2/\sigma^2)}{\sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)} \right.\right. \\
&\quad \left.\left. - \frac{\exp(-\|x_i - y\|^2/\sigma^2) \sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)}{\sum_{j=1}^{N} \exp(-\left\|x_j - y\right\|^2/\sigma^2) \sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)} \right) \right\| \\
&\leq \left\| \sum_{i=1}^{N} \hat{\phi}(x_i) \left(\frac{\exp(-\|x_i - z\|^2/\sigma^2)}{\sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)} \right.\right. \\
&\quad \left.\left. - \frac{\exp(-\|x_i - z\|^2/\sigma^2) \exp(-\delta^2/\sigma^2) \sum_{j=1}^{N} \exp(-\left\|x_j - y\right\|^2/\sigma^2) \exp(-\delta^2/\sigma^2)}{\sum_{j=1}^{N} \exp(-\left\|x_j - y\right\|^2/\sigma^2) \sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)} \right) \right\| \\
&= \left\| \sum_{i=1}^{N} \hat{\phi}(x_i) \frac{\exp(-\|x_i - z\|^2/\sigma^2)(1 - \exp(-2\delta^2/\sigma^2))}{\sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|^2/\sigma^2)} \right\| \\
&= \left\| \hat{\phi}^{smoothed}(z) \overbrace{1 - \exp(-2\delta^2)}^{\leq \delta} \right\| \\
&\leq \left\| \underbrace{\hat{\phi}^{smoothed}(z) \frac{1}{\sum_{j=1}^{N} \exp(-\left\|x_j - z\right\|/\sigma^2)}}_{L = max_x\left(\hat{\phi}^{smoothed}(x)\right)} \delta \right\|
\end{aligned}
$$

L exists as long as any Shapley value is bounded which follows from the bound on $f(\cdot)$.

The inequality $1-\exp(-2\delta^2) \leq \delta$ holds trivially for $\delta \geq 1$. For $0 < \delta < 1$, it holds that

$$\begin{aligned}
1-\exp(-2\delta^2) &< \delta \\
-2\delta^2 &> log(1-\delta) \\
-2\delta^2 &> -\sum_{k=1}^{\infty} \frac{\delta^k}{k} \\
\delta^2 &< \frac{1}{2}\sum_{k=1}^{\infty} \frac{\delta^k}{k} \\
1 &< \frac{1}{2\delta} + \frac{1}{4} + \frac{\delta}{6} < \frac{1}{2}\sum_{k=1}^{\infty} \frac{\delta^{k-2}}{k}
\end{aligned}$$

The last inequality follows from the fact that the roots of the polynomial $\frac{1}{2} + \frac{1}{4}\delta + \frac{\delta^2}{6}$ are outside of $(0,1)$.

A.8 Smoothed SHAP as Kernel Regressor

When smoothing is formulated as a kernel regressor, we can use this relationship to quantify the bias and the variance of the smoothed estimator from the true Shapley values building upon results for the multivariate Nadaraya-Watson estimator, as we will do in the following.

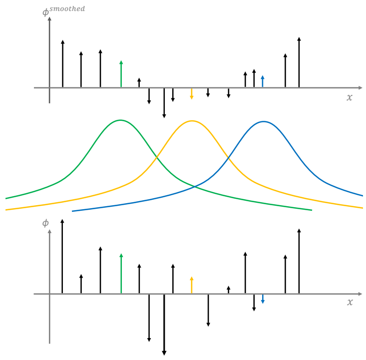
Assuming that $\hat{\phi}(j,x_i)$ for $i \in \{1,\ldots,N\}$ was computed as a mean over single-reference Shapley values A.1, that for each i the background data set has been sampled i.i.d., and that the black box f is bounded, then $\hat{\phi}(j,x_i)$ is a mean estimator and it follows from the Central Limit Theorem for a large number of references $L \to \infty$ that

$$\hat{\phi}(j,x_i) \approx \phi(j,x_i) + \varepsilon_i \tag{A.6}$$

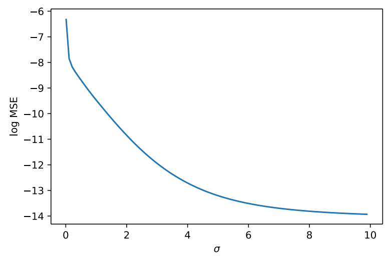
where $\varepsilon_i \sim \text{Normal}(0, \sigma_{ref}^2/L)$ denotes i.i.d. normal noise with σ_{ref}^2 being the variance of the single-reference Shapley values. Note that this form (with possibly different noise distribution) follows for each unbiased estimator $\hat{\phi}$ of ϕ. Since KernelSHAP is assumed to be biased, a bias-corrected version such as the one proposed by [218] can be used.

Let Σ denote the matrix of bandwidths, i.e. here $\Sigma = \text{diag}(\sigma_1,\ldots,\sigma_m)$. We then make the following assumptions

1. The Shapley estimator is unbiased, i.e. Equation A.6 holds.
2. The variance of $\text{Var}[\hat{\phi}(j,x) \mid x] =: \sigma_{\hat{\phi}}^2$, is continuous and positive.
3. The test instance x lies in the support of f. The probability density function of $\hat{\phi}(j,x)$, $p(\hat{\phi}(j,x))$, is continuously differentiable and bounded away from zero.



(a) The coloured exponential kernels illustrate how the smoothed Shapley values can be computed in a moving average like manner from the standard Shapley values.



(b) Log MSE for the black box $f(x) = x$ with $X \sim \text{Normal}(0,1)$ with 100 reference points for 1000 instances. The smoothed Shapley values were corrected by their offset $f(x) - \mathbb{E}_{p(X)}[f(X)]$.

Figure A4: Illustrations of Smoothed SHAP.

4. The kernel, i.e. density $d(x^*|x)$, is a symmetric and bounded PDF with finite second moment and square integrable, such as the exponential kernel.

5. The sequence of Σ is a deterministic sequence of positive definite symmetric matrices such that $N^{-1}|\Sigma|$ and each entry of Σ tend to 0 when $n \to \infty$.

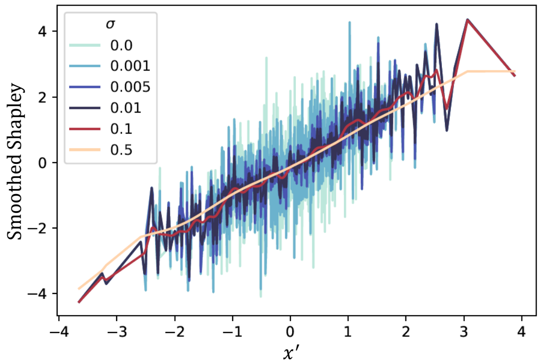
Note that the first two assumptions hold for Shapley values computed with KernelSHAP or the Shapley formula, as long as f has a finite expectation. It then holds that [78]

$$\begin{aligned} Bias[\hat{\phi}^{smoothed}(j,x), \phi(j,x)] &= \frac{1}{2}\mu_2(d)\,\mathrm{tr}\left(\Sigma \mathscr{H}_{\phi(j)}(x)\right) + o\left(\mathrm{tr}(\Sigma)\right) \\ \mathrm{Var}[\hat{\phi}^{smoothed}(j,x) \mid x_1, \ldots, x_N] &= \frac{1}{N\sigma_1 \cdot \ldots \sigma_m} \|d\|_2^2 \frac{\sigma_{\hat{\phi}}^2}{p(\hat{\phi}(j,x))} \left(1 + o(1)\right)) \end{aligned}$$

where $\mu_2 = \int uu^T d(u|0)du$, $\|d\|_2^2 = \int d(s|0)^2 ds$, $\mathscr{H}_f$ is the Hessian matrix of f, and $\mathrm{tr}(H)$ is the trace of the matrix H. For an exponential kernel it is $\|d\|_2^2 = 2^{-m}\pi^{-m/2}$. The Hessian $\mathscr{H}_{\phi(j)}$, as a measure of the complexity of the Shapley value function $\phi(j)(x) := \phi(j,x)$, in the bias term signalises that the bias increases the more instable Shapley values are to small perturbations of x. For smooth functions, i.e. linear functions, smoothing can indeed decrease the bias as seen in Figure A4b.

A.9 Modelling Measurement Error with Smoothed SHAP

As explained in the main part of the paper, Smoothed SHAP relates to modelling feature inclusion which allows to take the volatility in the measurement of the test instance into consideration.



	MSE
bike - daily average as measurement	
SHAP	291.8637±0.8304
Smoothed SHAP	**287.7668±0.8195**
bike - noisy measurement	
SHAP	195.0134±0.6174
Smoothed SHAP	**190.6908±0.6026**
adult - age groups	
SHAP	0.0037±0.0000
Smoothed SHAP	**0.0030±0.0000**

Table A2: MSE±standard error averaged over 5,000 runs on bike data. The bandwidth for the humidity feature of Smoothed SHAP was tuned on a validation set (20% of train data as described in Supplement A.11) over a range of 0.5 and 2. We see that Smoothed SHAP has always a significantly smaller MSE.

In a first simulated example, we assume a one-dimensional test instance x is measured with standard normal measurement error and linear black box, as presented in Figure A5. Since the measurement error is attributed to the only present feature, observations with close true value x' can have highly varying attributions because of the measurement error in x.

We will now illustrate Smoothed SHAP in the context of real-world examples. For simplicity we consider the prediction task with a boosted regressor on the bike data set with four features: `season`, `month`, `hour` and `humidity`. In a first scenario, we assume that we do not have access to the hourly humidity of the test instances, but only to their daily average. In a second case, we assume that the humidity measurement of the test instances is noisy, i.e. perturbed with normal noise of scale 0.1. Smoothed SHAP was then computed as a weighted average over the Shapley values of the training data and the SHAP values of the test data with measurement error. We set the bandwidths of the first three features close to 0 such that Smoothed SHAP of x' only depends on observations x that equal x' in the first three dimensions. The bandwidth of `hour` was tuned on a validation set. We also repeated a similar set up on the adult data set: a boosted classifier predicted high income based on the features `Workclass`, `Education-Num`, `Marital Status` and `Age`. In the test data, we do not have access to the age of an individual, but only to his age group (0-10, 10-30, 30-50, 50+). Again we only average over the SHAP values of individuals with the same first features as the test instance. Results are presented in Table A2.

A.10 Variance Analysis

Shapley values can be estimated by

$$\hat{\phi}(j,x) = \frac{1}{m} \sum_{S \subseteq \{1,\ldots,m\}/j} \left[\frac{1}{\binom{m-1}{|S|}} \left(\frac{1}{L} \sum_{i=1}^{L} f(x_{S \cup j}, x^*_{i,\overline{S}/j}) - \frac{1}{L} \sum_{i=1}^{L} f(x_S, x^*_{i,\overline{S}}) \right) \right].$$

We note that $\phi(j,x)$ is the expectation of a discrete variable $V \mid \{X^*_{i,\overline{S}/j}, X^*_{\overline{S}}\}_{S \subseteq \{1,\ldots,m\}/j}$ with

$$V \mid \{X^*_{\overline{S}/j}, X^*_{\overline{S}}\}_{S \subseteq \{1,\ldots,m\}/j} = \mathop{\mathbb{E}}_{r(X^*_{\overline{S}/j} \mid x_S)} \left[f(x_{S \cup j}, X^*_{\overline{S}/j}) \right] - \mathop{\mathbb{E}}_{r(X^*_{\overline{S}} \mid x_S)} \left[f(x_S, X^*_{\overline{S}}) \right]$$

with a support of size 2^{m-1} and probabilities $\left\{ \frac{1}{m\binom{m-1}{|S|}} \text{ for } S \subseteq \{1,\ldots,m\}/j \right\}$. Since the support of this discrete variable is estimated using a mean estimator, we can compute the variance of the estimated Shapley value as

$$\mathrm{Var}[\hat{\phi}(j)] = \frac{1}{m^2} \sum_{S \subseteq \{1,\ldots,m\}} \left[\frac{1}{\binom{m-1}{|S|}^2} \left(\frac{\sigma^2_{S \cup j}}{L} + \frac{\sigma^2_S}{L} \right) \right]$$

by independence and $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$ where σ^2_S is the variance of $f(x_S, X^*_{\overline{S}})$ which we estimate with the squared standard deviation of $\{f(x_S, x^*_{i,\overline{S}})\}_{i=1}^{L}$. For computing the variance of Neighbourhood SHAP, we use results from self-normalised importance sampling (see [219] or `http://statweb.stanford.edu/~owen/mc/Ch-var-is.pdf`) and estimate the variance σ^2_S now by

$$\hat{\sigma}^2_S = \frac{1}{(\sum_{k=1}^{L} d(x_S | x^*_{k,S}))^2} \sum_{i=1}^{L} d(x_S | x^*_{i,S})^2 (f(x_S, x^*_{i,\overline{S}}) - \bar{f}_w(x_S, x^*_{i,\overline{S}}))^2$$

where $\bar{f}_w(x_S, x^*_{i,\overline{S}})$ denotes the weighted average of $\{f(x_S, x^*_{i,\overline{S}})\}_{i=1}^{L}$ with weights $\left\{ \frac{d(x_S | x^*_{i,S})}{(\sum_{k=1}^{L} d(x_S | x^*_{k,S}))^2} \right\}_{i=1}^{L}$.
Since we do not know $p(\hat{\phi})$ to use the results from Supplement A.8, we instead approximate the variance of the smoothed Shapley values by

$$\mathrm{Var}[\hat{\phi}^{smoothed}(j,x)] = \frac{1}{(\sum_{i=1}^{L} w_i)^2} \sum_{S \subseteq \{1,\ldots,m\}} w_i^2 \sigma^2_{\hat{\phi}_i}$$

which follows again by independence and $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$ where $\sigma^2_{\hat{\phi}_i}$ is the variance of the estimator $\hat{\phi}(j, x_i)$.

Data Set	# Observations	# Features	URL
adult	32.561	11	https://archive.ics.uci.edu/ml/datasets/adult
bike	17.389	15	https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset
Boston	506	14	https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html
compas	6.172	12	https://github.com/dylan-slack/Fooling-LIME-SHAP/tree/master/data
mnist	60.000	784	https://pytorch.org/vision/stable/datasets.html

Table A3: Details on all real-world data sets used in the experiments.

Experiment	Type of Compute	Amount of Compute (in clock-time seconds)
Simulated experiment	2.4 GHz 6-Core Intel Core i5-9300H CPU	177
Adversarial attack	2.6 GHz 4-Core Intel Xeon Gold 6240 CPU	4,377
Lipschitz continuity	2.4 GHz 6-Core Intel Core i5-9300H CPU	$\approx$ 14,400 (correct up to hour)
adult	2,6 GHz 6-Core Intel Core i7	2,968
Boston	2,6 GHz 6-Core Intel Core i7	2,741
bike	2,6 GHz 6-Core Intel Core i7	427

Table A4: Total amount of compute for selected experiments.

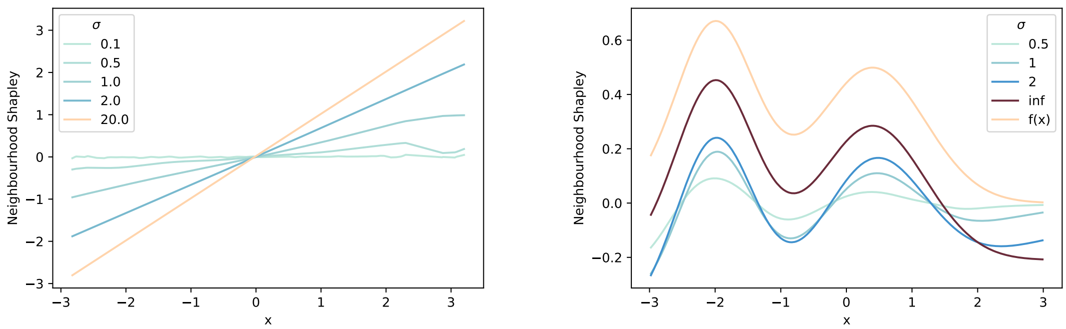
A.11 Experimental Details and Additional Experimental Results

In this section, we provide additional details on the experiments as presented in the main part, present complete results and add additional experiments which illustrate how our approaches perform.

A.11.1 Experimental Details

For our experiments, we used the code from the following public resources:

- KernelSHAP [57], available at
  `https://github.com/slundberg/shap` with MIT License
- Explanation GAME [71], available at
  `https://github.com/fiddler-labs/the-explanation-game-supplemental`
  (no license)
- LIME [22], available at
  `https://github.com/marcotcr/lime` with MIT License
- Fooling-LIME-SHAP [81], available at
  `https://github.com/dylan-slack/Fooling-LIME-SHAP` with MIT License
- MeLIME [68], available at
  `https://github.com/tiagobotari/melime` with MIT License
- Robust-interpret [89], available at
  `https://github.com/dmelis/robust_interpret` (no license)



(a) For $\sigma = 20$, the Neighbourhood Shapley values of $f(x) = x$ are visually not distinguishable from Shapley values. The smaller σ, the smaller the Shapley values since the expected model outcome within the neighbourhood of x gets closer to $f(x)$. Further, local Shapley values with small bandwidths are instable in the boundary areas where data is sparse.

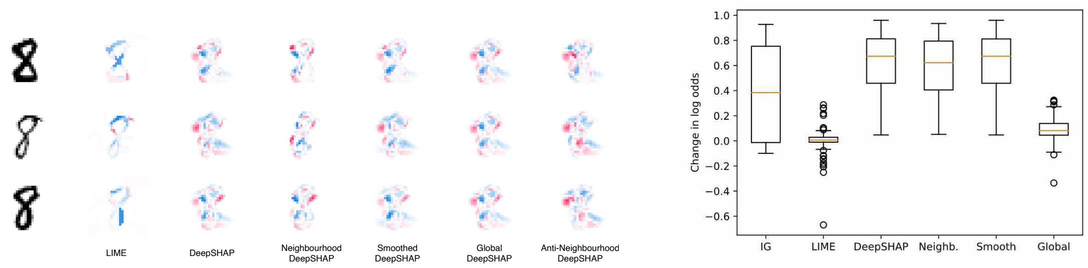
(b) Here the black box function is the CDF of the balanced mixture of $\text{Normal}(-2, 0.6)$ and $\text{Normal}(0.4, 1)$. The Shapley value attribution at $x = -1$ is positive, even though increasing x in a small neighbourhood would lead to a decrease in $f(x)$. The Neighbourhood Shapley values reflect this behaviour.

Figure A6: Neighbourhood Shapley values for 1000 different x sampled from $\text{Normal}(0, 1)$ and different bandwidths σ.

Black box models and other classification algorithms were trained using `sklearn` with default parameters. Please refer to Table A3 for details on the data sets. Data was splitted with a 80/20 split into train and test data randomly given the seeds, unless otherwise specified. All models were trained after data scaling with a standard scaler. We used the euclidean distance as a distance measure for continuous features. For categorical data, distance was set to one if the reference point had a different feature value, 0 otherwise. We used an exponential kernel as distance distribution $d(x^*|x)$. Following [71], we limited the visual inspection of the SHAP attributions on the UCI data sets presented in the main part of the paper in Figure 6 to the top 5 features which were selected using the Light GBM "split" feature importance, computed from the numbers of times the feature is used in a model.

A.11.2 Simulated Experiments

Simulated experiments comparing Neighbourhood SHAP and standard SHAP are included and explained in Figure A6. The local linear approximation presented in the main part and in Supplement A.1 were computed with a weighted Ridge Regression using the `LimeBASE` function of the LIME package. The data was sampled from the corresponding background data sets and the weights were computed with euclidean distance based exponential kernels.



(a) Attributions for the digit '3' of images trained on 100 background samples consisting of images of a 3 or an 8. We used 100 reference points and a bandwidth of $0.1 \cdot 784$ for the DeepSHAP alternatives. The Quickshift segmentation algorithm with a kernel size of 1, maximal distance of 1 and ratio of 0.95 was trained to obtain the highest possible number of segments (here 28) for the LIME algorithm.

(b) Change in Log odds when masking 100 randomly sampled images of an 8 to explain a 3, with a background data set of 100. We obtained a bandwidth of $0.25 \cdot 784$ for the Neighbourhood DeepSHAP approach and of $0.1 \cdot 784$ for the Smoothed DeepSHAP on a validation data set by optimising the bandwidth over a grid of [0.05, 0.1, ... 0.25, 0.3].

Figure A7: Predictive results on the MNIST data set.

A.11.3 Image Classification

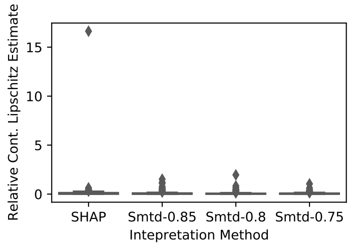
We applied the Neighbourhood and the Smoothed Shapley values also on the MNIST data set. We trained a convolutional network as in `https://github.com/slundberg/shap/blob/master/notebooks/image_examples/image_classification/PyTorch%20Deep%20Explainer%20MNIST%20example.ipynb` for ten epochs. We then applied DeepSHAP, Neighbourhood DeepSHAP, Smoothed DeepSHAP, Anti-Neighbourhood DeepSHAP (difference between DeepSHAP and Neighbourhood DeepSHAP). Refer to Figure A7 for the results.

In a second step, we mimicked the experimental set-up of [89] with results presented in Figure A8. We see that even though small amount of smoothing does not change the visual representation of the explanation in a detectable manner (Figure A7b), the Lipschitz estimates introduced by [89] decrease considerably.

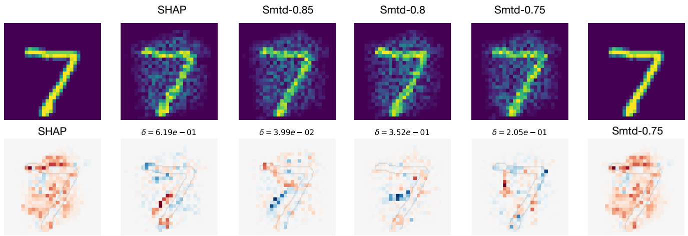
A.11.4 Adult Data

On the adult data we trained a LightGBM Classifier trained with `sklearn` with default parameters to predict an annual income of more than $50,000. When analysing the data set we consider an individual with the following features

- `capital gain` $= 0$



(a) Local Lipschitz estimates computed on 100 test points on the MNIST data set. We note that the Lipschitz estimates have a distribution with smaller outliers the smaller this threshold is, i.e. the more smoothing is used.



(b) Explanations of a CNN model prediction's on an example MNIST digit with Gaussian noise added to it. Here δ is the ratio $\|f(x) - f(x')\|_2 / \|x - x'\|_2$ for the perturbed x which was chosen such that it maximises the Lipschitz estimate of the explanation model on the test instance. The original digit image x' has been once explained with DeepSHAP and once with Smoothed DeepSHAP. We see that δ decreases with the amount of smoothing.

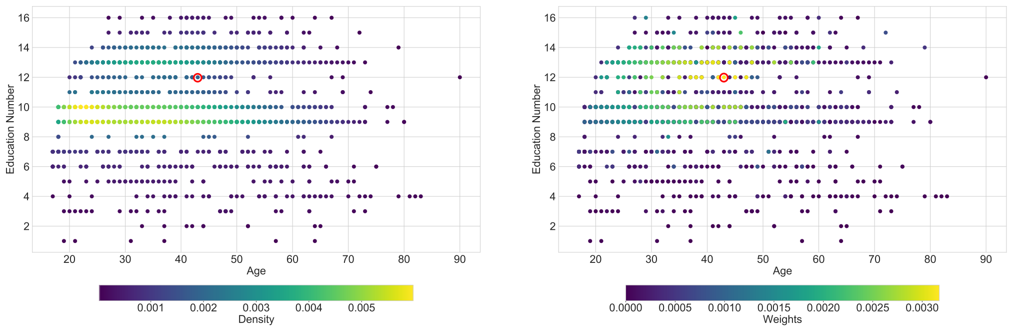
Figure A8: Results on Lipschitz continuity of SHAP and Smoothed SHAP on the MNIST data set. Smtd-0.8 denotes the Smoothed SHAP estimate where the bandwidth was chosen as large as possible such that the largest normalised weight is smaller than 0.8.

- `age` = 43
- `relationship` = Not-in-family
- `education number` = 12
- `marital-status` = Never-married
- `annual-income above $50,000` = True
- and a black box prediction of 0.094

The training set had 32561 instances whilst the test set had 16281 instances. Observations with missing values were deleted (N=3,620). The model fit was measured with the accuracy α. $\alpha = 14.3\%$ on the train set and $\alpha = 14.7\%$ on the test set for the parsimonious model. $\alpha = 11.0\%$ on the train set and $\alpha = 12.7\%$ on the test set for the entire model.
Following experimental results have been included in the following

- Descriptive plots for explaining attributions away from the centre of mass in Figure A9
- Bias and variance when using Smoothed SHAP in Figure A12a (please see Figure A27 for a plot of the bias of Smoothed SHAP over multiple tabular data sets.)
- A bar plot for comparing standard Shapley with Smoothed SHAP in Figure A12b
- Box plots to compare the variance across methods in Figure A10
- Plots for varying kernel widths for the parsimonious model in Figure A11

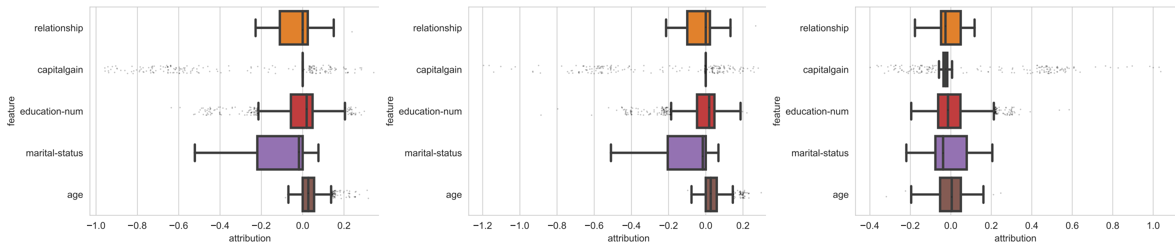
- Neighbourhood SHAP for varying kernel widths: Entire model in Figure A13
- Smoothed SHAP results across background data set in Figure A14



(a) Coloured by kernel-density estimates using Gaussian kernels.

(b) Coloured by Neighbourhood SHAP weights for the individual of interest. Kernel width was equal to 2.

Figure A9: Scatter plot of the Education Number by Age across reference points. Individual of interest lies in the red circle.

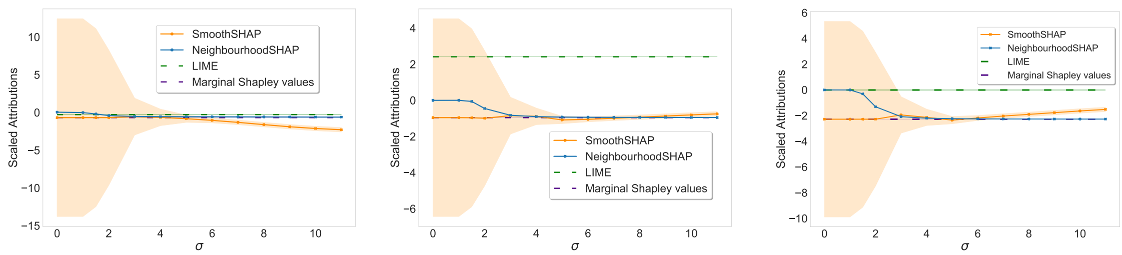


(a) Standard Shapley values

(b) Neighbourhood SHAP

(c) Smoothed SHAP

Figure A10: Comparison of box plots on the parsimonious model: standard Shapley values, Neighbourhood SHAP, Smoothed SHAP. Kernel width equal to 5.

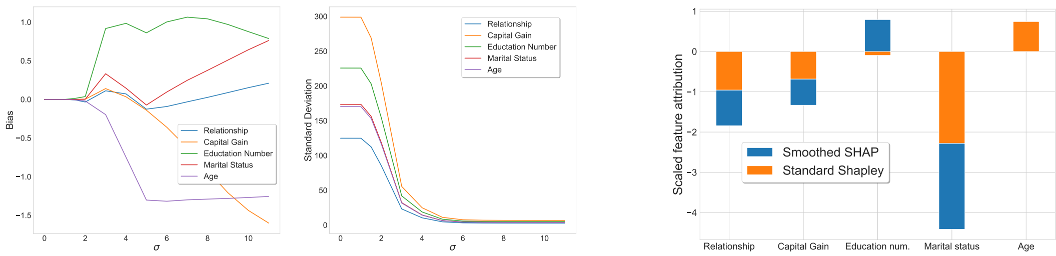


(a) Capital Gain

(b) Relationship

(c) Marital Status

Figure A11: Scaled attributions at a given test instance for varying kernel widths computed with 2000 reference points in the adult data sets. Bounds for LIME have been computed over 2000 runs, while the Shapley bounds have been estimated with their theoretical formula as outlined in Supplement A.10.



(a) Bias and standard deviation for Smoothed SHAP values of our individual of interest by kernel widths, using 2000 reference points.

(b) Comparison of standard Shapley values and Smoothed SHAP with a kernel width equal to 2, using 2000 reference points.

Figure A12

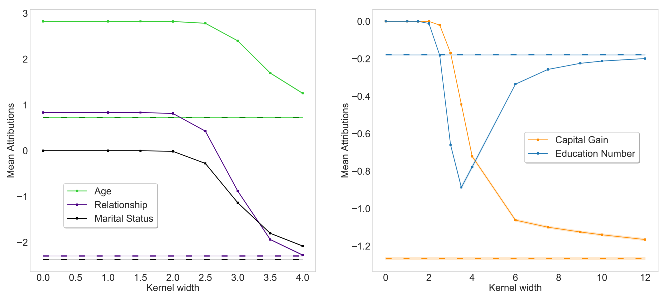


Figure A13: Mean attributions of Neighbourhood for varying kernel widths, with our model trained on all 14 features. Only features from the parsimonious model are represented in this plot. Values are computed for our individual of interest, using 2000 reference points. Marginal Shapley values correspond to dashed lines; Neighbourhood SHAP values correspond to solid lines.

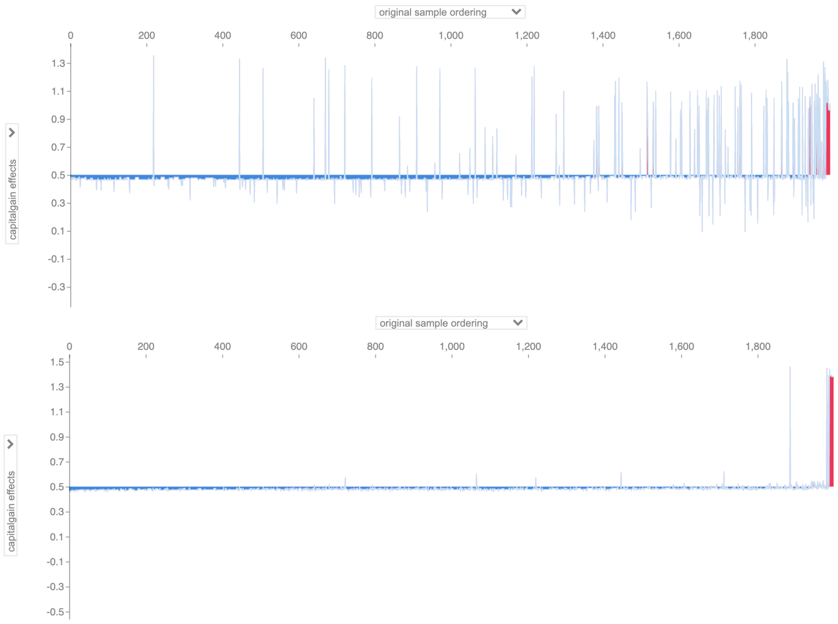


Figure A14: Comparison of the marginal Shapley values (top) and Smoothed SHAP with kernel width 5 (bottom) for computing the effects of Relationship in the Adult Income data set using the parsimonious model. Feature attributions are sorted by similarity according to a preliminary PCA analysis across a subset of 2000 samples from the Adult Income data set, using 2000 reference points. Plot was created using the KernelSHAP package [57].

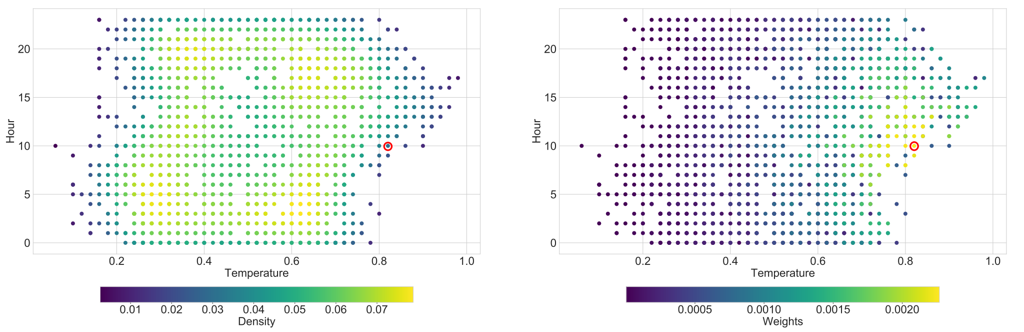
A.11.5 Bike Data

On the bike data we trained a LightGBM Regressor trained with `sklearn` with default parameters to predict the hourly number of bike rentals. The hour we analyse in the bike data is

- `workingday` = True
- `hour` = 10
- `temperature` = 0.82
- `humidity` = 0.56
- `season` = 3
- `hourly number of bike rentals` = 218
- and prediction of 110.7

The training set had 8645 instances whilst the test set had 8734 instances. The model fit was measured with the coefficient of determination. $R^2 = 0.934$ on the train set and $R^2 = 0.632$ on the test set for the parsimonious model. $R^2 = 0.959$ on the train set and $R^2 = 0.641$ on the test set for the entire model. Similarly to before we have included following results for the bike data set:

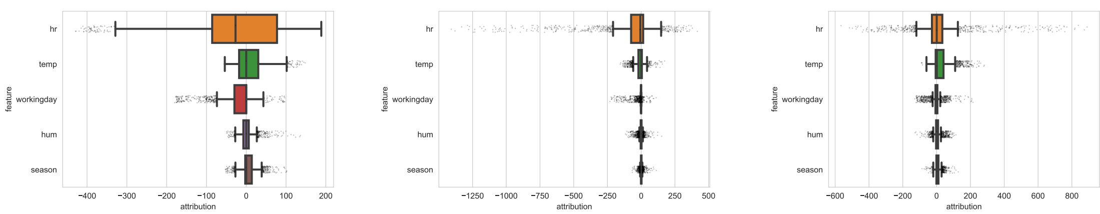
- Descriptive plots for explaining attributions away from the center of mass in Figure A15
- Bias and variance when using Smoothed SHAP in Figure A18a
- A bar plot for comparing standard Shapley with SmoothedSHAP in Figure A18b
- Box plots to compare the variance across methods in Figure A16
- Plots for varying kernel widths for the parsimonious model in Figure A17
- Neighbourhood SHAP for varying kernel widths: Entire model in Figure A20
- Smoothed SHAP results across background data set in Figure A19



(a) Coloured by kernel-density estimates using Gaussian kernels.

(b) Coloured by Neighbourhood SHAP weights for the individual of interest. Kernel width was equal to 2.

Figure A15: Scatter plot of `hour` by `Temperature` across reference points. Individual of interest lies in the red circle.

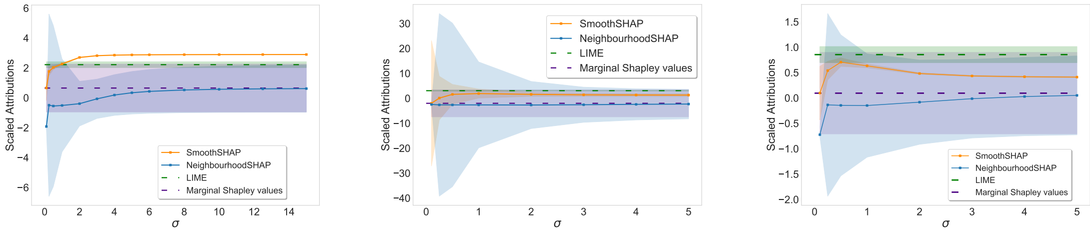


(a) Standard Shapley values

(b) Neighbourhood SHAP

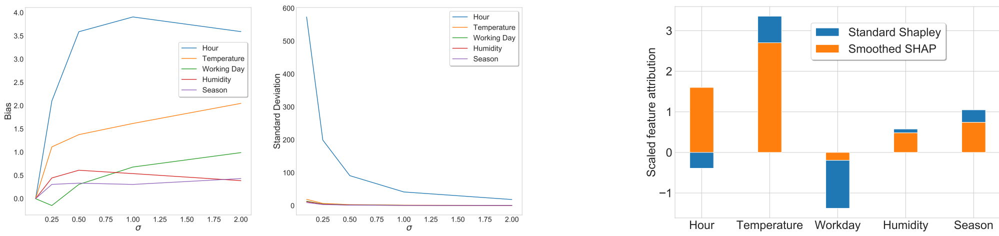
(c) Smoothed SHAP

Figure A16: Comparison of box plots on the parsimonious model: standard Shapley values, Neighbourhood SHAP, Smoothed SHAP. Kernel width equal to 2.



(a) Temperature (b) Hour (c) Humidity

Figure A17: Scaled attributions at a given test instance for varying kernel widths computed with 2000 reference points in the adult data sets. Bounds for LIME have been computed over 2000 runs, while the Shapley bounds have been estimated with their theoretical formula as outlined in Supplement A.10.



(a) Bias and standard deviation for Smoothed SHAP values of our individual of interest by kernel widths, using 2000 reference points.

(b) Comparison of standard Shapley values and Smoothed SHAP with a kernel width equal to 2, using 2000 reference points.

Figure A18

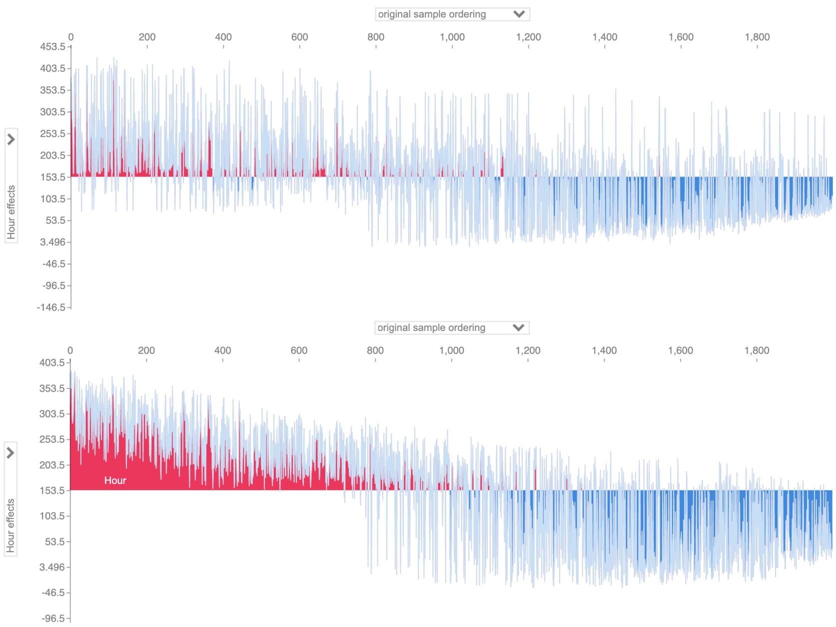


Figure A19: Comparison of the marginal Shapley values (top) and Smoothed SHAP with kernel width 2 (bottom) for computing the effects of `hour` in the bike data set using the parsimonious model. Feature attributions are sorted by similarity according to a preliminary PCA analysis across a subset of 2000 samples from the Adult Income data set, using 2000 reference points. Plot was created using the KernelSHAP package [57].

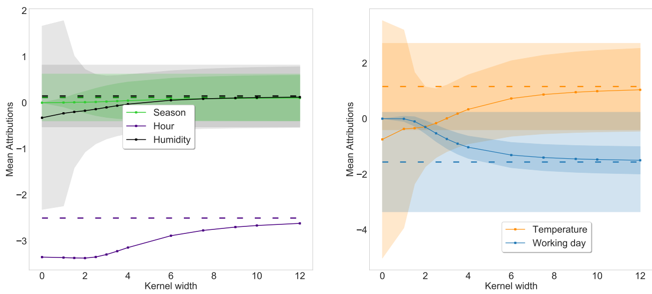


Figure A20: Mean attributions of Neighbourhood SHAP for varying kernel widths, with our model trained on all 10 features. Only features from the parsimonious model are represented in this plot. Values are computed for the individual of interest, using 2000 reference points. Marginal Shapley values correspond to dashed lines; Neighbourhood SHAP values correspond to solid lines. For `hour`, the standard error is not represented as it is too wide: approximately 128 for the standard Shapley value, while it decreases from approximately 582 std to 131 for the Neighbourhood Shapley values.

A.11.6 Boston Housing Data

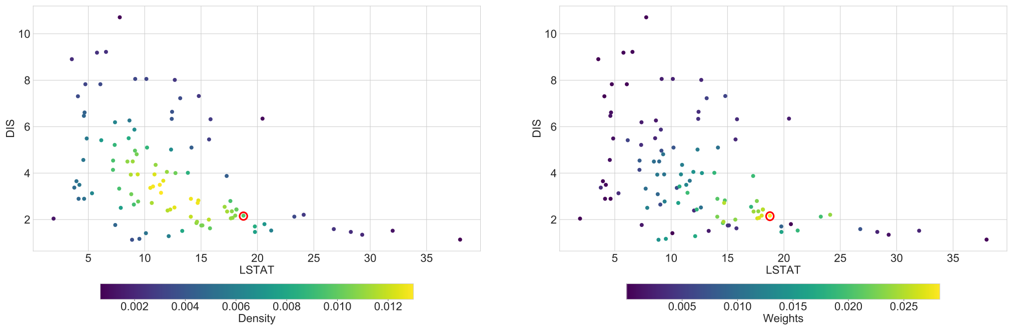
On the Boston housing data we trained a LightGBM Regressor with `sklearn` with default parameters to predict the median value of owner-occupied homes in $1000's. The dwelling analysed has following properties

- `Percentage of lower status of the population (LSTAT)` = 18.8
- `Average number of rooms per dwelling (RM)` = 6.05
- `Per capita crime rate by town (CRIM)` = 5.29
- `Weighted distances to five Boston employment centres (DIS)` = 2.17
- `Pupil-teacher ratio by town (PTRATIO)` = 20.2
- `Median value in $1000's` = 23.2
- and prediction of 16.4

The training set had 404 instances whilst the test set had 102 instances. The model fit was measured with the coefficient of determination. $R^2 = 0.963$ on the train set and $R^2 = 0.661$ on the test set for the parsimonious model. $R^2 = 0.977$ on the train set and $R^2 = 0.699$ on the test set for the entire model.

- Descriptive plots for explaining attributions away from the center of mass in Figure A21
- Bias and variance when using Smoothed SHAP in Figure A24a
- A bar plot for comparing standard Shapley with SmoothedSHAP in Figure A24b
- Box plots to compare the variance across methods in Figure A22

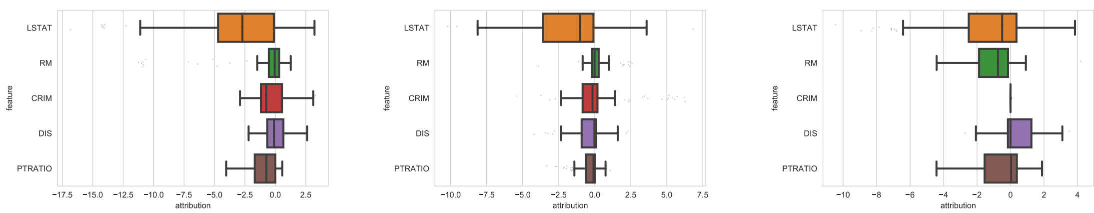
- Plots for varying kernel widths for the parsimonious model in Figure A23
- Neighbourhood SHAP for varying kernel widths: Entire model in Figure A25
- Smoothed SHAP results across background data set in Figure A26



(a) Coloured by kernel-density estimates using Gaussian kernels.

(b) Coloured by Neighbourhood SHAP weights for the individual of interest. Kernel width was equal to 2.

Figure A21: Scatter plot of the Percentage of lower status population (LSTAT) by Weighted distances to five Boston employment centres (DIS) across reference points. Individual of interest lies in the red circle.

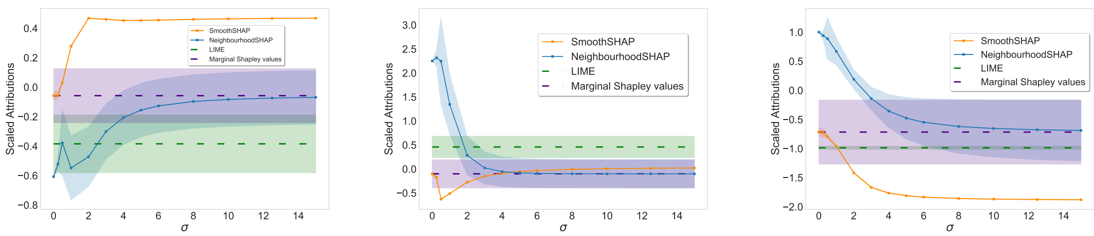


(a) Standard Shapley values

(b) Neighbourhood SHAP

(c) Smoothed SHAP

Figure A22: Comparison of box plots on the parsimonious model: standard Shapley values, Neighbourhood SHAP, Smoothed SHAP. Kernel width equal to 2.

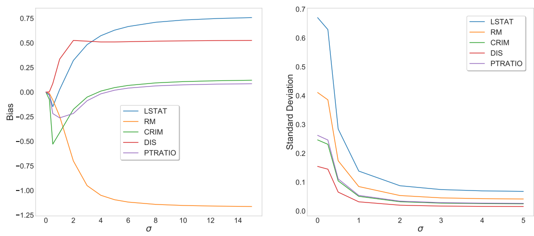


(a) DIS

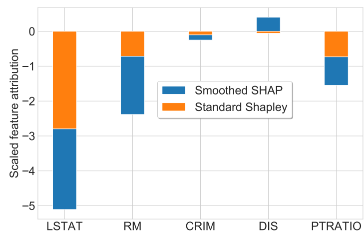
(b) CRIM

(c) RM

Figure A23: Scaled attributions at a given test instance for varying kernel widths computed with 102 reference points in the adult data sets. Bounds for LIME have been computed over 2000 runs, while the Shapley bounds have been estimated with their theoretical formula as outlined in Supplement A.10.



(a) Bias and standard deviation for Smoothed SHAP values of our individual of interest by kernel widths, using 102 reference points.



(b) Comparison of standard Shapley values and Smoothed SHAP with a kernel width equal to 2, using 102 reference points.

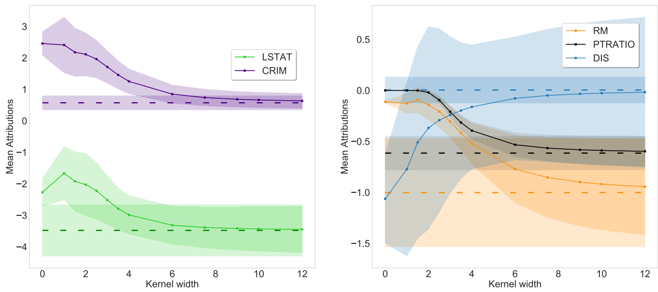


Figure A25: Mean attributions of Neighbourhood SHAP for varying kernel widths, with our model trained on all 10 features. Only features from the parsimonious model are represented in this plot. Values are computed for our individual of interest, using 102 reference points. Marginal Shapley values correspond to dashed lines; Neighbourhood SHAP values correspond to solid lines.

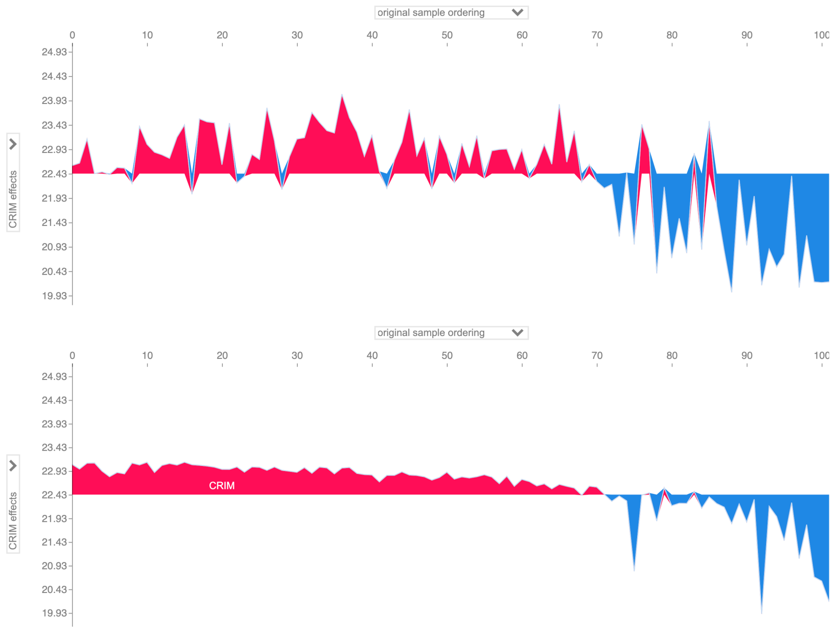


Figure A26: Comparison of the marginal Shapley values (top) and Smoothed SHAP with kernel width 2 (bottom) for computing the effects of CRIM in the bike data set using the parsimonious model. Feature attributions are sorted by similarity according to a preliminary PCA analysis across a subset of 102 samples from the Adult Income data set, using 2000 reference points. Plot was created using the KernelSHAP package [57].

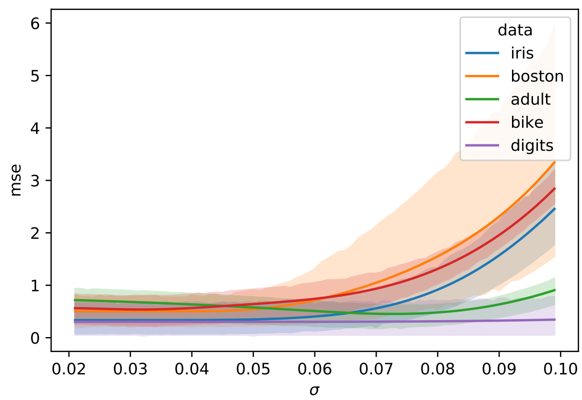


Figure A27: Bias of Smoothed SHAP for different kernel widths. While most data sets reach their minimum MSE at a positive kernel width (bike: 0.03, adult: 0.08, Boston: 0.03, digits: 0.1, iris: 0), this minimum is insignificant.

A.11.7 ROAR metric

	Adult (α)		**Bike (R^2)**		**Boston (R^2)**	
	Local mean	**Global mean**	**Local mean**	**Global mean**	**Local mean**	**Global mean**
SHAP	0.276	0.098	0.232	0.189	0.349	0.154
Neighbourhood SHAP	0.364	0.271	0.259	0.312	0.389	0.320
Smoothed SHAP	0.233	0.280	0.391	0.306	0.177	0.399

Table A5: Absolute change in evaluation metrics (R^2 or accuracy) on the test data when imputing with local and global mean. Abbreviations: α=accuracy

Appendix B

Supplementary Materials: PWSHAP

B.1 Causal Inference Background

Confounder A confounder is a variable that is associated with both the exposure and the outcome, causing a spurious correlation. For instance, summer is associated with eating ice cream and getting sunburns, but there is no causal relationship between the two.

Mediator A mediator is a variable that is both an effect of the exposure and a cause of the outcome. In presence of a mediator, the total effect can be broken into two parts: the direct and indirect effect.

Moderator A mediator is a pre-exposure variable for which the causal effect is heterogeneous in subgroups.

Propensity score model A propensity score model is a function that predicts exposure from the observed covariates. We note it $\pi^*(c) = P(T = 1|C = c)$ and note π an estimate of π^*.

Potential outcome As defined by the Rubin causal model [109], a potential outcome $Y(t)$ is the value that Y would take if T were set by (hypothetical) intervention to the value t.

Identification assumptions

- **No interference** For a given individual i, this assumption implies that $Y_i(t)$ represents the value that Y would have taken for individual i if T had been set to t for individual i, i.e the potential value of Y_i if T_i had been set to t.

- **Consistency** For a given individual i, $T_i = t \Rightarrow Y_i = Y_i(t)$. This means that for individuals who actually received exposure level t, their observed outcome is the same as what it would have been had they received exposure level t via an hypothetical intervention.

- **Conditional exchangeability** For a given individual i, we assume that conditional on C, the actual exposure level T is independent of each of the potential outcomes: $Y(t) \perp T \mid \mathbf{C}, \forall t$

Average Treatment Effect (ATE) The Average Treatment Effect for a binary treatment is the average difference in potential outcomes: $\mathbb{E}[Y(1) - Y(0)]$.
Conditional Average Treatment Effect (CATE) The Conditional Average Treatment Effect for a binary treatment, conditioned on C is the average difference in potential outcomes: $\mathbb{E}[Y(1) - Y(0)|C = c]$. If C is a sufficient adjustment set, i.e. conditional exchangeability w.r.t. C holds then the CATE can be identified as $\mathbb{E}[Y|T = 1, C = c] - \mathbb{E}[Y|T = 0, C = c]$.

Controlled Direct Effect Let $Y(t,m)$ be the potential outcome under exposure level $T = t$ and mediator level $M = m$. The controlled direct effect of T on outcome Y comparing $T = t$ with $T = t^*$ and setting M to m measures the effect of T on Y not mediated through M i.e. the effect of T on Y after intervening to fix the mediator to some value m. The controlled direct effect for individual i is then $\mathrm{CDE}_i(t,t^*,m) = Y_i(t,m) - Y_i(t^*,m)$ [220]

Natural Direct Effect The natural direct effect is defined as the difference between the value of the counterfactual outcome if the individual were exposed to $T = t$ and the value of the counterfactual outcome if the same individual were instead exposed to $T = t^*$, with the mediator M taking whatever value it would have taken at the reference value of the exposure $T = t^*$: $Y(t,M(t^*))$˘$Y(t^*,M(t^*))$ [220]

Natural Indirect Effect The natural indirect effect is the difference, having set the exposure to a fixed level $T = t$, between the value of the counterfactual outcome if the mediator M took whatever value it would have taken at a level of the exposure $T = t$ and the value of the counterfactual outcome if the mediator assumed whatever value it would have taken at a reference level of the exposure $T = t^*$: $Y(t,M(t))$˘$Y(t,M(t^*))$ [220]

B.2 Notations

- Coalition-wise Shapley *values* $\phi^f_{S,T}$ are the individual terms for a coalition in the weighted sum of the original definition of Shapley value, hence the term, *value*. It is common to use ϕ -even if specific to a coalition S- in reference to Shapley values.

- Coalition-wise Shapley *effects* $\Psi_{T\to Y|C_S}$ are the causal *effects* identified in coalition-wise Shapley *values* after dividing by the propensity weights. We use Ψ for effects, and describe the effect of T on Y along the multiple paths through the covariates C_S. We symbolise this by using the subscript $T \to Y|C_S$.

- Path-wise Shapley *effects* Ψ_{C_i} are similar to coalition-wise *effects* since they are also obtained after dividing by the weights. However, the conditioning is only on the features on a single causal pathway. We thus still use Ψ as it is an effect, but show that the conditioning only bears upon a path using the subscript C_i.

B.3 Further Related Work

B.3.1 Shapley Values: Definitions

Off-manifold Shapley values

$$\phi^f_j(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [\mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}})] - \mathbb{E}[f(x_S, X^*_{\overline{S}})]]$$

On-manifold Shapley values

$$\begin{aligned}\phi^f_j(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} &[\mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid X_{S\cup\{j\}} = x_{S\cup\{j\}}] \\ &- \mathbb{E}[f(x_S, X^*_{\overline{S}})] \mid X_S = x_S]\end{aligned}$$

Causal Shapley values

$$\begin{aligned}\phi^f_j(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} &[\mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] \\ &- \mathbb{E}[f(x_S, X^*_{\overline{S}})] \mid \mathrm{do}(X_S = x_S)]\end{aligned}$$

where the contribution of feature j, $\phi^f_{j,S}(x)$, is decomposed into a direct and an indirect effect as follows:

$$\begin{aligned}
\phi^f_{j,S}(x) &= \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}}) - \mathbb{E}[f(x_S, X^*_{\bar{S}})] \mid \mathrm{do}(X_S = x_S)] \\
&= \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_S = x_S)] - \mathbb{E}[f(x_S, X^*_{\bar{S}})] \mid \mathrm{do}(X_S = x_S)] \\
&+ \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] \\
&- \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_S = x_S)]
\end{aligned}$$

where, in the last equality, the first line is the direct effect and the second line the indirect effect. See Section B.3.3 for details. Note that "interventional" Shapley values in Equation 3 of [61] are not causal Shapley values but actually off-manifold Shapley values : the do-operator is written in Equation 3, but it is misleading as the do-operator is said to intervene "by breaking the dependence between features in [the coalition] and the remaining features", i.e. making the expectation marginal as in off-manifold Shapley values.

B.3.2 Shapley Values: Axioms

In this section, in line with Section 4.1, the Shapley value $\phi^f_j(x)$ is always taken with respect to value function v, unless specified otherwise. In the latter case, if the value function is v', then the Shapley value is noted $\phi^f_j(x; v')$ instead. Shapley values have been shown to satisfy the four following axioms.

Dummy: A feature j receives a zero attribution if it has no possible contribution, i.e. $v_f(S \cup \{j\}, x) = v_f(S, x)$ for all $S \subseteq \{1, \ldots, m\}$.

Symmetry: Two features that always have the same contribution receive equal attribution, i.e. $v_f(S \cup \{i\}, x) = v(S \cup \{j\}, x)$ for all S not containing i or j then $\phi^f_i(x) = \phi^f_j(x)$.

Efficiency: The attributions of all features sum to the total value of all features. Formally, $\sum_j \phi^f_j(x) = v_f(\{1, .., m\}, x)$.

Linearity: For any value function v that is a linear combination of two other value functions u and w (i.e. $v(S) = \alpha u(S) + \beta w(S)$), the Shapley values of v are equal to the corresponding linear combination of the Shapley values of u and w
(i.e. $\phi^f_i(x; v) = \alpha \phi^f_i(x; u) + \beta \phi^f_i(x; w)\Big)$.

B.3.3 Causal Shapley Values

Heskes et. al introduced the Causal Shapley values in 2020 [104]. For a coalition S, the contribution of feature j $\phi^f_{j,S}(x)$ is decomposed into a direct and an indirect effect:

$$
\begin{aligned}
&\phi^f_{j,S}(x) \\
&= \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}}) - \mathbb{E}[f(x_S, X^*_{\bar{S}})] \mid \mathrm{do}(X_S = x_S)] \\
&= \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_S = x_S)] - \mathbb{E}[f(x_S, X^*_{\bar{S}})] \mid \mathrm{do}(X_S = x_S)] \\
&+ \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] - \mathbb{E}[f(x_{S\cup\{j\}}, X^*_{\overline{S\cup\{j\}}}) \mid \mathrm{do}(X_S = x_S)]
\end{aligned}
$$

where, in the last equality, the first line is the direct effect and the second line the indirect effect. The direct effect measures the expected change in prediction when the stochastic feature X_j is replaced by its feature value x_j, without changing the distribution of the other 'out-of-coalition' features. The indirect effect measures the difference in expectation when the distribution of the other 'out-of-coalition' features changes due to the additional intervention $do\left(X_j = x_j\right)$. The direct and indirect parts of Shapley values are then be computed by taking a, possibly weighted, average over all coalitions.
We note that in the problem setup of Section 4.3.1 , if all covariates are pre-treatment then under mild assumptions the indirect effect of the treatment will be zero, as outlined in the following Proposition.

Property 6 (Indirect part of Causal Shapley). Let S be a coalition containing pre-treatment covariates only. We assume that an unobserved (latent) variable generates all pre-treatment covariates. Then the indirect part of the Causal Shapley values of an exposure is null, i.e. we have

$$\mathbb{E}\left[f\left(C_{\bar{S}}, c_S, t\right) \mid \mathrm{do}(C_S = c_S, T = t)\right] - \mathbb{E}\left[f\left(C_{\bar{S}}, c_S, t\right) \mid \mathrm{do}(C_S = c_S)\right] = 0. \tag{B.1}$$

The proof can be found in Supplement B.10.7.

B.3.4 Edge-Based/Flow-Based Approaches to Shapley Values

Given that the proposed approach is model-agnostic, in this paragraph we will not review model-specific approaches that are considered to be "bespoke in nature and do not solve the problem of explainability in general" [82]. Similarly, we do not review methods that violate *implementation invariance*[1]. Pan et. al [221] leverage Shapley values computation to define a new quantity that distributes credit for model disparity amongst the paths in a causal graph. However, the resulting quantity isn't a Shapley value itself. Shapley Flow

[1]Implementation invariance imposes that two black-box models that compute the same mathematical function have identical attributions for all features, regardless of how being implemented differently.

(SF) assigns credits to "sink-to-node" paths. To do so, SF only considers orderings that are consistent with a depth first search. Furthermore, SF modifies the original definition of Shapley values by only explaining within successive cuts of the graphs or "explanation boundaries". Such cuts are considered as alternative models to be explained. Given this modification, it is unclear what connection the explanations generated by SF exhibit with the overall model. Recursive Shapley [111] is an edge-based approach which only considers active edges. Although it provides useful insights for mediation analysis, this method overlooks the impact of confounders. Ultimately, unlike our approach both Shapley Flow and Recursive Shapley aren't additive methods, but instead hold the property of "flow conservation" which allows a parent node to split its credit amongst its children. In contrast, the *efficiency* axiom of Shapley values ensure that the attributions of all features sum up to the model outcome $f(x)$. We argue that Shapley efficiency is more relevant in a regression setting, whereas flow conservation should be used for analysis of data with intrinsic ordering.

B.3.5 Other Causal Approaches to Interpreting Black-Box Models

Explaining black-box models in a causal manner remains challenging to this day. Zhao et. al [222] expand on the use of partial dependence plot (PDP), where the dependence on a set of covariates is computed by taking the expectation of the model over the marginal distribution of all other covariates. They note that the PDP formula is similar to Pearl's back-door adjustment. More specifically, marrying Shapley values with causal reasoning has been an active research question. Janzing et. al [70] considers the model's prediction process itself to be a causal process: from features to model inputs and ultimately model output. The authors claim that marginal Shapley values can be apprehended in terms of *do*-calculus, if we consider that setting a feature to a given value is equivalent to intervening on it. As such, marginal or so-called off-manifold Shapley values may be sufficient to explain that specific causal process but this approach is contrived as it does not consider the real-world causal relationships between features. This approach however does not acknowledge any underlying real-world causal structure. New causality based formulations of Shapley values have been proposed to compute feature attributions from a hypothesised causal structure of the data. Asymmetric Shapley values [82] use conditioning by observation but only consider causally-consistent coalitions i.e coalitions such that known causal ancestors precede their descendants. The resulting explanations quantify the impact a given feature has on model prediction while its descendants remain unspecified. As a result, they ignore downstream effects in favour of root causes [110].

Below is a table showing a summary comparison of existing causality-based or graph-based Shapley approaches with PWSHAP. Node efficiency refers to the original efficiency property of Shapley values: $f(x) = \phi_0^f(x) + \sum_{i=1}^M \phi_i^f(x)$, where $\phi_j^f(x)$ is the contribution of feature j to $f(x)$ and $\phi_0^f(x) = \mathbb{E}[f(X))]$ is the averaged prediction with the expectation over the observed data distribution. By game at each node/within each boundary we describe the fact that Shapley Flow and Recursive Shapley consider successive cuts from the graphs. Flow conservation or cut efficiency is the equivalent efficiency property, within such a cut. We refer the reader to the corresponding papers for further details.

	Shapley Flow	Recursive Shapley	Asymmetric Shapley	Causal Shapley	PWSHAP
Flow conservation or cut efficiency	X	X			
Node efficiency			X	X	X
Node-based			X	X	X
Edge-based		X			
Source-to-sink path-based	X				
Path-based					X
Game at each node or within each boundary of explanation	X	X			
Ignores direct effects			X		
Fidelity to original Shapley			X (but violates symmetry axiom)	X	X

Table C1: Comparison between PWSAP and existing causality-based or graph-based Shapley approaches

Ultimately, in a previous work [223], Sani et. al (2020) introduce an XAI method which uses auxiliary interpretable labels that are assumed to be readily available. Although their method has shown to perform well, it relies on the latter strong assumption which limits its applicability. Note that, like in PWSHAP, this approach further requires external information (besides the model and input/output data). We believe Sani et. al's method can be complementary to ours. For instance, in settings where the causal quantity of interest is identified from the obtained Partial Ancestral Graph, one may train a model regressing predictions on the interpretable labels and apply PWSHAP to the resulting model. We thank our anonymous reviewers for this remark.

B.3.6 Further discussion on an XAI model's dependency to the DAG

We view the fact that our approach is agnostic to the choice of (compatible) DAG as a strength, as it allows different experts to explain the black-box model output according to their own causal beliefs about the data or phenomenon being studied. For example, take the classical DAG with confounders, a treatment and an outcome (in Figure 4.2, graph 2). The black-box model takes all confounders and the treatment as inputs. However it does not "know" whether the covariates are actually confounders or mediators, and generates its output regardless. In PWSHAP, given that we have the additional (external) knowledge of the DAG representing our beliefs, we would interpret the output as a conditional average treatment effect. More generally, a model only gives indications about causal quantities that are already identified with statistical quantities according to prior causal assumptions, as in the causal roadmap by Petersen and van der Laan (2014) [224]). Although causal discovery using both the model outputs *and the data* can yield a set of candidate DAGs that share the same edges, in the form of a Partial Ancestral Graph (PAG), some directions may be missing and one may still need to choose a DAG amongst multiple alternatives.

B.3.7 Bias, Mediation and Moderation Analysis

Our approach has added value compared to existing methods for sensitivity, mediation and moderation analysis. In the following paragraph, we review the state of the art with regards to each of these objectives.

Moderation analysis A common approach to assess moderation is to (i) fit an Heterogeneous Treatment Effect (HTE) model that predicts an individual's treatment effect from a set of covariates (ii) find subgroups with similar treatment effects. Subgroups can be infered directly from the data [114, 115], from the individual predicted treatment effects [225] or using statistical hypothesis tests [40, 116, 226]. The main drawback of subgroup findings methods is that they are prone under-powered and time-consuming. Holmes et. al [116] introduce a partitioning method which controls the type I error, however it is still limited to comparing subgroups two by two. Another approach to moderation analysis is to use interpretable models to predict HTEs. Nilsson et. al [227] build two potential outcomes models (treated/untreated) and fit a regression model to predict their difference from the covariates of interest. Regression coefficients are ultimately used as a measure of moderation, but this solution is prone to model misspecification. Explanation methods, and in particular feature attribution models allow for a finer-grained understanding of the sources of heterogeneity. More recently, Wu et. al suggested to use Distillation to generate explanations of HTE models and assess moderation induced by each covariate [228]. However, explanation models such as Distillation that involve building a simpler surrogate model have received criticism for *approximating* the target black-box function instead of explaining it [28]. Ultimately, HTE models are built without taking the causal structure into consideration the causal structure e.g. not conditioning on post-treatment features. To the best of our knowledge, our approach is the first method that can assign attribution to moderators directly from the outcome regression model whilst acknowledging the posited causal structure and the rules of *do*-calculus by Pearl.

Sensitivity analysis In most treatment effect estimation studies, it is assumed that all confounders of treatment and outcome are observed. This is a strong assumption and one might wonder whether results will be greatly perturbed or not by the presence of an unobserved confounder. Sensitivity analysis generally aims at determining what the impact of a given amount of unobserved confounding would be on causal conclusions of the study. In particular, as we have no access to the unobserved confounder, we make assumptions about its relationship with treatment and outcome. One line of work is to assume parameters for this relationship and infer the rest of the model when values of these parameters are fixed, e.g. via maximum likelihood [112, 229, 230]. As a result, one can check the change in treatment effects with fixed values of the unobserved confounder [229] or draw contour plots showing the bias depending on the parameters [112, 230]. Another line of work assumes a fixed ratio between the propensity score with only observed covariates and a variation of the propensity score that also includes unobserved confounders [231, 232, 233]. This ratio quantifies how much hidden confounding is present (with a ratio of 1 being no hidden confounding) and is set by the user. As a result, one can deduce intervals for inference quantities like p-values or treatment effects from a given ratio. This can be leveraged to find the lowest ratio that makes the interval reach thresholds invalidating causal conclusions, e.g. 0.05 for a p-value or zero for the treatment effects. The higher the ratio has to be, the more robust to unobserved confounding the study. This idea is close to the E-value, a scalar metric representing the minimal amount of unobserved confounding needed to fully explain away the treatment-outcone relationship [234]. Although sensitivity analysis can also be applied to assess the role of a given observed confounder [112, 230], it remains different from our local bias analysis approach that only applies to observed confounders. However, unlike most sensitivity analysis methods, our approach does not rely on parameters or assumptions other than the joint distribution of the data, and it summarises the confounding of a given pre-treatment covariate in a single bias scalar.

Mediation analysis The state-of-the-art approach to mediation analysis is based on Natural Direct and Indirect Effects [220]. Computation of Natural Direct/Indirect Effects requires four assumptions [220] : 1) no unmeasured confounding for the exposure-outcome relationship, 2) no unmeasured confounding for the mediator-outcome relationship, 3) no unmeasured confounding for the exposure-mediator relationship, 4) no mediator-outcome confounding that is itself affected by the exposure. These are strong hypotheses, with the latter typically being considered to be unrealistic. PWSHAP also implicitly relies on these assumptions. However, it is common to assume unconfoundedness of the exposure-outcome relationship. Indeed, the exposure is naturally or experimentally randomised in many mediation analysis settings, such as fairness studies (e.g. when sex or race are the exposure). Thus, without confounders, PWSHAP effects are causal in these settings. Further, our approach to mediation analysis is more faithful to causal inference than Causal Shapley. By comparing CDEs, we assess the effect of setting the given mediator to its value. By contrast, Causal Shapley breaks the relationship between treatment and mediator when intervening on the mediator in the indirect effect. A local mediation analysis example is given in Supplement B.4.3. Supplement B.5 details the application of PWSHAP to dependent mediators or in presence of both confounders and mediators.

B.4 Detailed Results for the "Building Blocks" Examples

B.4.1 Local Moderation Analysis

In the following, we prove the result of the first example where treatment is randomised

We assume the following model

$$\begin{aligned} C_1, C_2 &\sim \text{Uniform}(0,1), \quad C_1 \perp\!\!\!\perp C_2 \\ Y &= \beta T + \gamma_1 C_1 + \gamma_2 C_2 + \alpha_1 T C_1 + \alpha_2 T C_2 + \varepsilon \end{aligned}$$

with $\mathbb{E}[\varepsilon | T, C_1, C_2] = 0$. Assuming the true outcome model is known, our aim is to explain the following black-box: $f^*(c_1, c_2, t) = \beta t + \gamma_1 c_1 + \gamma_2 c_2 + \alpha_1 t c_1 + \alpha_2 t c_2$. We further assume that treatment is randomised by taking $T \sim \text{Bernoulli}(p)$.

We note that, from Property 1,

$$\begin{aligned} \phi_{T,\{C_1,C_2\}}^{f^*,\text{obs}}(c_1,c_2,t) = \phi_{T,\{C_1,C_2\}}^{f^*,\text{causal}}(c_1,c_2,t) &= (t-p)(\beta + \alpha_1 c_1 + \alpha_2 c_2) \\ \phi_{T,\{C_1\}}^{f^*,\text{obs}}(c_1,t) = \phi_{T,\{C_1\}}^{f^*,\text{causal}}(c_1,t) &= (t-p)(\beta + \alpha_1 c_1 + \alpha_2/2) \\ \phi_{T,\{C_2\}}^{f^*,\text{obs}}(c_2,t) = \phi_{T,\{C_2\}}^{f^*,\text{causal}}(c_2,t) &= (t-p)(\beta + \alpha_1/2 + \alpha_2 c_2) \\ \phi_{T,\emptyset}^{f^*,\text{obs}}(t) = \phi_{T,\emptyset}^{f^*,\text{causal}}(t) &= (t-p)(\beta + \alpha_1/2 + \alpha_2/2) \end{aligned}$$

All of these Causal Shapley values only correspond to direct effects, as the indirect effect is zero from Property 6.

B.4.2 Local Bias Analysis

We compare PWSHAP with Causal Shapley on a bias analysis example where C_1 and C_2 are distributed as before, but we assume instead that treatment allocation depends only on one covariate C_1: $\mathbb{E}[T|C_1,C_2] = C_1^\alpha$ which implies that C_1 is both a confounder and a moderator whereas C_2 only acts as a moderator. PWSHAP effects are then given as:

$$\Psi^{f^*}_{\substack{T\leftarrow C_1\rightarrow Y,\\ C_1:T\rightarrow Y}} = \alpha_1(c_1 - \frac{\alpha+1}{\alpha+2}) - \gamma_1\frac{\alpha+1}{2(\alpha+2)} \qquad \Psi^{f^*}_{C_2:T\rightarrow Y} = \alpha_2(c_2 - \frac{1}{2})$$
$$\Psi^{f^*}_{T\rightarrow Y|\emptyset} = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \frac{\alpha_2}{2}.$$

We note that $\mathbb{E}[\Psi^{f^*}_{C_2:T\rightarrow Y}(C_1,C_2)] = 0$ but $\mathbb{E}[\Psi^{f^*}_{\substack{T\leftarrow C_1\rightarrow Y,\\ C_1:T\rightarrow Y}}(C_1,C_2)] \neq 0$. This illustrates not only Lemma 1, but also to the following corollary where we do not assume the model is true.

Corollary 1 (Integration of the local confounding effect, black-box model). Let C_1, C_2 be two pre-treatment covariates such that $C_2 \perp\!\!\!\perp T|C_1$. Then the integral of the local confounding effect w.r.t. C_2 on the joint distribution of covariates is null

$$\mathbb{E}[\Psi^{f}_{T\leftarrow C_2\rightarrow Y}(C_1,C_2)] = 0.$$

The proof can be found in Supplement B.10.8. By comparison, in Causal Shapley values, only the direct part is non null:

$$\phi^{f^*,\text{causal}}_{T,\text{direct}} = \beta(t - \frac{c_1^\alpha}{2} - \frac{1}{2(\alpha+1)}) + \alpha_1(\frac{c_1}{2}(t-c_1^\alpha) + \frac{1}{2}(\frac{t}{2} - \frac{1}{\alpha+2}))$$
$$+\alpha_2((\frac{c_2}{3} + \frac{1}{12})(t-c_1^\alpha) + (\frac{c_2}{6} + \frac{1}{6}(t - \frac{1}{\alpha+1})).$$

Proof : First let's note that

$$\mathbb{E}[Y|T=1,C_1=c_1,C_2=c_2] - \mathbb{E}[Y|T=0,C_1=c_1,C_2=c_2] = \beta + \alpha_1 c_1 + \alpha_2 a_2$$

Then we show that :

$$\mathbb{E}[Y|T=1,C_1=c_1] - \mathbb{E}[Y|T=0,C_1=c_1] = \beta + \alpha_1 c_1 + \frac{a_2}{2}$$
$$\mathbb{E}[Y|T=1,C_2=c_2] - \mathbb{E}[Y|T=0,C_2=c_2] = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \alpha_2 c_2$$
$$\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \frac{\alpha_2}{2}$$

First, let us note that, using independence of C_1 and C_2,

$$\mathbb{E}[T|C_1=c_1] = \mathbb{E}[\mathbb{E}[T|C_1=c_1,C_2]|C_1=c_1] = \mathbb{E}[c_1^\alpha|C_1=c_1] = c_1^\alpha$$

$$\mathbb{E}[T|C_2 = c_2] = \mathbb{E}[\mathbb{E}[T|C_2 = c_2, C_1]|C_2 = c_2] = \mathbb{E}[C_1^\alpha|C_2 = c_2] = \frac{1}{\alpha+1}$$
$$\mathbb{E}[T] = \frac{1}{\alpha+1}$$

By Bayes's rule and independence of C_1 and C_2,

$$p(c_2|c_1, t=1) = \frac{p(t=1|c_1,c_2)p(c_1)p(c_2)}{p(t=1|c_1)p(c_1)} = \frac{p(t=1|c_1,c_2)}{p(t=1|c_1)} = \frac{c_1^\alpha}{c_1^\alpha} = 1$$

and, similarly,

$$p(c_2|c_1, t=0) = 1$$

As a result,

$$\mathbb{E}[C_2|c_1, t=1] = \mathbb{E}[C_2|c_1, t=0] = \frac{1}{2}$$
$$\mathbb{E}[C_2|c_1, t=1] - \mathbb{E}[C_2|c_1, t=0] = 0$$

and

$$\begin{aligned}
&\mathbb{E}[Y|T=1, C_1 = c_1] - \mathbb{E}[Y|T=0, C_1 = c_1] \\
&= \beta + \gamma_2(\mathbb{E}[C_2|c_1, t=1] - \mathbb{E}[C_2|c_1, t=0] + \alpha_1 c_1 + \alpha_2 \mathbb{E}[C_2|c_1, t=1] \\
&= \beta + \alpha_1 c_1 + \frac{a_2}{2}
\end{aligned}$$

which proves the first equality. For the second equality, we have

$$p(c_1|c_2, t=1) = \frac{p(t=1|c_1,c_2)p(c_1)p(c_2)}{p(t=1|c_2)p(c_2)} = \frac{c_1^\alpha}{\frac{1}{\alpha+1}} = (\alpha+1)c_1^\alpha$$

and, similarly,

$$p(c_1|c_2, t=0) = \frac{1-c_1^\alpha}{1-\frac{1}{\alpha+1}}$$

Thereby, we obtain

$$\mathbb{E}[C_1|c_2, t=1] = \frac{\alpha+1}{\alpha+2}$$
$$\mathbb{E}[C_1|c_2, t=0] = \frac{\alpha+1}{2(\alpha+2)}$$
$$\mathbb{E}[C_1|c_2, t=1] - \mathbb{E}[C_1|c_2, t=0] = \frac{\alpha+1}{2(\alpha+2)}$$

and

$$\mathbb{E}[Y|T=1, C_2 = c_2] - \mathbb{E}[Y|T=0, C_1 = c_1]$$

$$
\begin{aligned}
&= \beta + \gamma_1(\mathbb{E}[C_1|c_2, t=1] - \mathbb{E}[C_1|c_2, t=0]) \\
&\qquad + \alpha_2 c_2 + \alpha_1 \mathbb{E}[C_1|c_2, t=1] \\
&= \beta + \gamma_1 \frac{\alpha+1}{2(\alpha+2)} + \alpha_2 c_2 + \alpha_1 \frac{\alpha+1}{\alpha+2}
\end{aligned}
$$

Similarly, for the third equality, we note that, as before,

$$
\begin{aligned}
p(c_2|t=1) &= 1 \\
p(c_2|t=0) &= 1 \\
p(c_1|t=1) &= (\alpha+1)c_1^{\alpha} \\
p(c_1|t=0) &= \frac{1-c_1^{\alpha}}{1-\frac{1}{\alpha+1}}
\end{aligned}
$$

which leads to, as before,

$$
\begin{aligned}
\mathbb{E}[C_2|t=1] &= \frac{1}{2} \\
\mathbb{E}[C_2|t=0] &= \frac{1}{2} \\
\mathbb{E}[C_2|t=1] - \mathbb{E}[C_2|c_1, t=0] &= 0 \\
\mathbb{E}[C_1|t=1] &= \frac{\alpha+1}{\alpha+2} \\
\mathbb{E}[C_1|t=0] &= \frac{\alpha+1}{2(\alpha+2)} \\
\mathbb{E}[C_1|t=1] - \mathbb{E}[C_1|c_2, t=0] &= \frac{\alpha+1}{2(\alpha+2)}
\end{aligned}
$$

thereby

$$
\begin{aligned}
\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] &= \beta + \gamma_1(\mathbb{E}[C_1|T=1] - \mathbb{E}[C_1|T=0]) \\
&\quad + \gamma_2(\mathbb{E}[C_2|T=1] - \mathbb{E}[C_2|T=0]) \\
&\quad + \alpha_1 \mathbb{E}[C_1|T=1] + \alpha_2 \mathbb{E}[C_2|T=1] \\
&= \beta + \gamma_1 \frac{\alpha+1}{2(\alpha+2)} + \alpha_1 \frac{\alpha+1}{\alpha+2} + \frac{a_2}{2}
\end{aligned}
$$

Now, for Causal Shapley values, we can show that

$$
\begin{aligned}
\mathbb{E}[Y|\mathrm{do}(t, c_1, c_2)] - \mathbb{E}[Y|\mathrm{do}(c_1, c_2)] &= (t - c_1^{\alpha})(\beta + \alpha_1 c_1 + \alpha_2 c_2) \\
\mathbb{E}[Y|\mathrm{do}(t, c_1)] - \mathbb{E}[Y|\mathrm{do}(c_1)] &= \beta(t - c_1^{\alpha}) + \alpha_1 c_1 (t - c_1^{\alpha}) + \alpha_2 (\frac{t}{2} - \frac{c_1^{\alpha}}{2}) \\
\mathbb{E}[Y|\mathrm{do}(t, c_2)] - \mathbb{E}[Y|\mathrm{do}(c_2)] &= \beta(t - \frac{1}{\alpha+1}) + \alpha_2 c_2 (t - \frac{1}{\alpha+1}) + \alpha_1 (\frac{t}{2} - \frac{1}{\alpha+2}) \\
\mathbb{E}[Y|\mathrm{do}(t)] - \mathbb{E}[Y] &= \beta(t - \frac{1}{\alpha+1}) + \alpha_1 (\frac{t}{2} - \frac{1}{\alpha+2}) + \alpha_2 (\frac{t}{2} - \frac{1}{2(\alpha+1)})
\end{aligned}
$$

as $\mathbb{E}[TC_1|c_1] = c_1^{\alpha+1}$, $\mathbb{E}[TC_1|c_2] = \mathbb{E}[TC_1] = \frac{1}{\alpha+2}$, $\mathbb{E}[TC_2|c_1] = \frac{c_1^\alpha}{2}$, $\mathbb{E}[TC_2|c_2] = \frac{c_2}{\alpha+1}$, $\mathbb{E}[TC_2] = \frac{1}{2(\alpha+1)}$

B.4.3 Local Mediation Analysis

We compare PWSHAP with Causal Shapley on a mediation analysis example inspired by the Berkeley dataset [235]. An algorithm predicts the probability of success of an applicant to a college. In this example, $X = (T, Q, D)$ where T is the gender of the applicant Q is an exam result and D is the department. We assume $Q \sim \text{Uniform}(0,1)$, $D|T = 0 \sim \text{Bernoulli}(0.8)$ and $D|T = 1 \sim \text{Bernoulli}(0.2)$. D is a mediator of gender however Q is only an ancestor of the outcome, and not a mediator. Our black-box is the true outcome model, and $Y = \alpha_Q Q + \alpha_D D + \alpha_T T + \alpha_{DT} DT + \alpha_{QT} QT + \varepsilon$, with $\mathbb{E}[\varepsilon|D, Q, T] = 0$.

$$\Psi^{f^*}_{T\to D\to Y} = 0.6\alpha_D + \alpha_{DT}(d - \frac{1}{5}) \qquad \Psi^{f^*}_{T\to Q\to Y} = \alpha_{QT}(q - \frac{1}{2})$$

$$\begin{aligned}
\Psi^{f^*}_{T\to Y|\emptyset} &= \alpha_T - 0.6\alpha_D + \frac{\alpha_{DT}}{5} + \frac{\alpha_{QT}}{2} \\
\phi^{f^*,\text{causal}}_{T,\text{direct}} &= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}[\frac{d}{2}(t - \frac{1}{2}) + \frac{1}{2}(\frac{t}{2} - \frac{1}{10})] + \frac{\alpha_{QT}}{2}(t - \frac{1}{2})(q + \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\text{indirect}} &= \frac{\alpha_D}{2}(\frac{3}{10} - \frac{3t}{5})
\end{aligned}$$

We note that $\int_q \Psi^{f^*}_{T\to Q\to Y}(p,q)dp(q) = 0$ but $\int_d \Psi^{f^*}_{T\to D\to Y}(p,q)dp(d) \neq 0$.
Proof : Coalition-wise Shapley effects are :

$$\begin{aligned}
\Psi^{f^*}_{T\to Y|D,Q}(d,q) &= \text{CDE}(d,q) \\
&= \mathbb{E}[Y|T = 1, D = d, Q = q] - \mathbb{E}[Y|T = 0, D = d, Q = q] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT} q \\
\Psi^{f^*}_{T\to Y|D}(d) &= \text{CDE}(d) \\
&= \mathbb{E}[Y|T = 1, D = d] - \mathbb{E}[Y|T = 0, D = d] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT}\mathbb{E}[Q|T = 1] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT}\mathbb{E}[Q] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT}\frac{1}{2} \\
\Psi^{f^*}_{T\to Y|Q}(q) &= \text{CDE}(q) \\
&= \mathbb{E}[Y|T = 1, Q = q] - \mathbb{E}[Y|T = 0, Q = q]
\end{aligned}$$

$$
\begin{aligned}
&= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT}q + \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{\alpha_{DT}}{5} + \alpha_{QT}q - \alpha_D\frac{3}{5} \\
\Psi^{f^*}_{T\to Y|\emptyset}(q) &= \text{ATE} \\
&= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \\
&= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT}\mathbb{E}[Q|T=1] \\
&+ \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{\alpha_{DT}}{5} + \frac{\alpha_{QT}}{2} - \alpha_D\frac{3}{5}
\end{aligned}
$$

As a result, we deduce path-wise effects

$$
\begin{aligned}
\Psi^{f^*}_{T\to D\to Y} &= \Psi^{f^*}_{T\to Y|D,Q} - \Psi^{f^*}_{T\to Y|Q} = 0.6\alpha_D + \alpha_{DT}(d - \frac{1}{5}) \\
\Psi^{f^*}_{T\to Q\to Y} &= \Psi^{f^*}_{T\to Y|D,Q} - \Psi^{f^*}_{T\to Y|D} = \alpha_{QT}(q - \frac{1}{2})
\end{aligned}
$$

Causal Shapley values are :

$$
\begin{aligned}
\phi^{f^*,\text{causal}}_{T,\{D,Q\},\text{direct}}(d,q,t) &= (\alpha_T + \alpha_{DT}d + \alpha_{QT}q)(t - \frac{1}{2}) \\
&= \mathbb{E}[f(d,q,t)|\text{do}(d,q)] - \mathbb{E}[f(d,q,T)|\text{do}(d,q)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(dt - d\mathbb{E}[T]) + \alpha_{QT}(qt - q\mathbb{E}[T]) \\
&= (t - \frac{1}{2})(\alpha_T + \alpha_{DT}d + \alpha_{QT}q) \\
\phi^{f^*,\text{causal}}_{T,\{D,Q\},\text{indirect}}(d,q,t) &= \mathbb{E}[f(d,q,t)|\text{do}(d,q,t)] - \mathbb{E}[f(d,q,t)|\text{do}(d,q)] \\
&= 0 \\
\phi^{f^*,\text{causal}}_{T,\{D\},\text{direct}}(d,t) &= \mathbb{E}[f(d,Q,t)|\text{do}(d)] - \mathbb{E}[f(d,Q,T)|\text{do}(d)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(dt - d\mathbb{E}[T]) + \alpha_{QT}(\mathbb{E}[Q]t - \mathbb{E}[QT]) \\
&= (\alpha_T + \alpha_{DT}d + \frac{\alpha_{QT}}{2})(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\{D\},\text{indirect}}(d,t) &= \mathbb{E}[f(d,Q,t)|\text{do}(d,t)] - \mathbb{E}[f(d,Q,t)|\text{do}(d)] \\
&= 0 \\
\phi^{f^*,\text{causal}}_{T,\{Q\},\text{direct}}(q,t) &= \mathbb{E}[f(D,q,t)|\text{do}(q)] - \mathbb{E}[f(D,q,T)|\text{do}(q)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(\mathbb{E}[D]t - \mathbb{E}[DT]) + \alpha_{QT}(qt - q\mathbb{E}[T]) \\
&= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}(\frac{t}{2} - \frac{1}{10}) + \alpha_{QT}q(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\{Q\},\text{indirect}}(q,t)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}[f(D,q,t)|\mathrm{do}(q,t)] - \mathbb{E}[f(D,q,t)|\mathrm{do}(q)] \\
&= \alpha_D(\mathbb{E}[D|t] - \mathbb{E}[D]) \\
&= \alpha_D(\frac{3}{10} - \frac{3t}{5}) \\
\phi^{f^*,\text{causal}}_{T,\emptyset,\text{direct}}(t) &= \mathbb{E}[f(D,Q,t)] - \mathbb{E}[f(D,Q,T)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(\mathbb{E}[D]t - \mathbb{E}[DT]) \\
&\quad + \alpha_{QT}(\mathbb{E}[Q]t - q\mathbb{E}[QT]) \\
&= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}(\frac{t}{2} - \frac{1}{10}) + \frac{\alpha_{QT}}{2}(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\emptyset,\text{indirect}}(t) & \\
&= \mathbb{E}[f(D,Q,t)|\mathrm{do}(t)] - \mathbb{E}[f(D,Q,t)] \\
&= \alpha_D(\mathbb{E}[D|t] - \mathbb{E}[D]) \\
&= \alpha_D(\frac{3}{10} - \frac{3t}{5})
\end{aligned}
$$

Summing them altogether with appropriate binomial weights, we have

$$
\begin{aligned}
\phi^{f^*,\text{causal}}_{T,\text{direct}} &= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}[\frac{d}{2}(t - \frac{1}{2}) + \frac{1}{2}(\frac{t}{2} - \frac{1}{10})] + \frac{\alpha_{QT}}{2}(t - \frac{1}{2})(q + \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\text{indirect}} &= \frac{\alpha_D}{2}(\frac{3}{10} - \frac{3t}{5})
\end{aligned}
$$

B.5 Further Results for Complex "Building Blocks" DAGS

B.5.1 A Mix of Confounders and Mediators

Property 7 (Ancestors of outcome with confounders). Let M_1, M_2 be post-treatment and pre-outcome variables. and assume that covariates C include all confounders of the relationships between T, Y and (M_1, M_2). Further assume that $M_2 \perp\!\!\!\perp T, M_1 | C$; that is, M_2 is not really a mediator because $M_2 \perp\!\!\!\perp T | C$, and is independent of M_1 conditionally on T, C. Then, for any value c of confounders C and m_1 of M_1

$$\mathbb{E}[\Psi^{f^*}_{T\to M_2 \to Y}(c, m_1, M_2) \mid C = c] = 0$$

We now assume the model :

$$
\begin{aligned}
C_1 &\sim \text{Bernoulli}(\frac{1}{2}) \\
C_2 &\sim \text{Bernoulli}(\frac{1}{2})
\end{aligned}
$$

$$\begin{aligned}
Q|C_1 &\sim \text{Bernoulli}(1-C_1)\\
T|C_1 &\sim \text{Bernoulli}(C_1)\\
D|T,C_1 &\sim \text{Bernoulli}(\frac{4}{5}-\frac{3}{5}\frac{T+C_1}{2})\\
Y &= \alpha_Q Q+\alpha_D D+\alpha_T T+\alpha_1 C_1+\alpha_2 C_2+\alpha_{DT}DT+\alpha_{QT}QT\\
&\quad +\alpha_{1T}C_1T+\alpha_{2T}C_2T+\varepsilon,\ \text{with } \mathbb{E}[\varepsilon|D,Q,T]=0.
\end{aligned}$$

We have,

$$\begin{aligned}
\Psi^{f^*}_{T\to Y|C_1,C_2,D,Q}(c_1,c_2,d,q) &= \alpha_T+\alpha_{DT}d+\alpha_{QT}q+\alpha_{1T}c_1+\alpha_{2T}c_2\\
\Psi^{f^*}_{T\to Y|C_1,C_2,Q}(c_1,c_2,q) &= \alpha_T+\alpha_{DT}\mathbb{E}[D|T=1,c_1,c_2,q]+\alpha_{QT}q+\alpha_{1T}c_1+\alpha_{2T}c_2\\
&\quad +\alpha_D(\mathbb{E}[D|T=1,c_1,c_2,q]-\mathbb{E}[D|T=0,c_1,c_2,q])\\
&= \alpha_T+\alpha_{DT}(\frac{1}{2}-\frac{3c_1}{10})+\alpha_{QT}q+\alpha_{1T}c_1+\alpha_{2T}c_2-\frac{3\alpha_D}{10}\\
\Psi^{f^*}_{T\to Y|C_1,C_2,D}(c_1,c_2,d) &= \alpha_T+\alpha_{DT}d+\alpha_{QT}\mathbb{E}[D|T=1,c_1,c_2,d]+\alpha_{1T}c_1+\alpha_{2T}c_2\\
&\quad +\alpha_Q(\mathbb{E}[Q|T=1,c_1,c_2,d]-\mathbb{E}[Q|T=0,c_1,c_2,d])\\
&= \alpha_T+\alpha_{DT}d+\alpha_{QT}(1-c_1)+\alpha_{1T}c_1+\alpha_{2T}c_2\\
\Psi^{f^*}_{T\to Y|C_1,D,Q}(c_1,d,q) &= \alpha_T+\alpha_{DT}d+\alpha_{QT}q+\alpha_{1T}c_1+\frac{\alpha_{2T}}{2}\\
\Psi^{f^*}_{T\to Y|C_2,D,Q}(c_2,d,q) &= \alpha_T+\alpha_{DT}d+\alpha_{QT}q+\alpha_{2T}c_2+\alpha_{1T}\mathbb{E}[C_1|T=1,d,q]\\
&\quad +\alpha_1(\mathbb{E}[C_1|T=1,d,q]-\mathbb{E}[C_1|T=0,d,q])\\
&\qquad \text{with, in the general case, } \mathbb{E}[C_1|T=t,d,q]\neq 0\ \ \forall t\\
&\qquad \text{and } \mathbb{E}[C_1|T=1,d,q]-\mathbb{E}[C_1|T=0,d,q]\neq 0
\end{aligned}$$

Thereby,

1. Local mediating effects are given as

   $$\Psi^{f^*}_Q(c_1,c_2,d,q)=\alpha_{QT}(q-(1-c_1))\ \text{ and }\ \Psi^{f^*}_D(c_1,c_2,d,q)=\frac{3\alpha_D}{10}+\alpha_{DT}(\frac{3c_1}{10}-\frac{1}{2})$$

   so $\mathbb{E}[\Psi^{f^*}_Q(c_1,c_2,d,Q)|c_1,c_2]=0$ but $\mathbb{E}[\Psi^{f^*}_D(c_1,c_2,D,q)|c_1,c_2]\neq 0$. This illustrates the relevance of Property 5 to isolate the fact that Q is not an actual mediator conditionally on confounders.

2. Local confounding effects are given as

   $$\begin{aligned}
   \Psi^{f^*}_{C_2}(c_1,c_2,d,q) &= \alpha_{2T}(c_2-\frac{1}{2})\\
   \Psi^{f^*}_{C_1}(c_1,c_2,d,q) &= \alpha_1(\mathbb{E}[C_1|T=0,d,q]-\mathbb{E}[C_1|T=1,d,q])-\alpha_{1T}\mathbb{E}[C_1|T=1,d,q]
   \end{aligned}$$

so $\mathbb{E}[\Psi^{f^*}_{C_2}(C_1,C_2,d,q)] = 0$ but $\mathbb{E}[\Psi^{f^*}_{C_1}(C_1,C_2,D,q)] \neq 0$. This illustrates the relevance of the two following results, which themselves generalise results of Section 4.6.1, to isolate the fact that C_2 is not an actual confounder of the relationship between treatment-outcome, treatment-mediator and mediator-outcome relationships.

Lemma 3 (Integration of the local confounding effect with mediators, true model). Let M denote post-treatment and pre-outcome variables. Define $\mathscr{H}(C)$ as follows :

$$\mathscr{H}(C): \ \forall t,m,\ Y(t,m) \perp\!\!\!\perp T|C \text{ and } Y(t,m) \perp\!\!\!\perp M|T,C,$$

or, in other words, C includes all confounders of the treatment-outcome and mediator-outcome relationships. We further assume consistency of the potential outcome, i.e. $Y(T,M) = Y$. Let C_1,C_2 be two pre-treatment covariates such that $\mathscr{H}(C_1,C_2)$ holds. If, additionally, C_2 is not a confounder, i.e. $\mathscr{H}(C_1)$ holds, then the integral of the local confounding effect of f^* w.r.t. C_2 on the joint distribution of covariates for fixed values of mediators is null, i.e.

$$\forall m,\ \mathbb{E}[\Psi^{f^*}_{C_2}(C_1,C_2,m)] = 0. \tag{B.2}$$

Corollary 2 (Integration of the local confounding effect with mediators, black-box model). Let C_1,C_2 be two pre-treatment covariates and M post-treatment and pre-outcome variables such that $C_2 \perp\!\!\!\perp T,M|C_1$. Then the integral of the local confounding effect w.r.t. C_2 on the joint distribution of covariates is null, i.e.

$$\forall m,\ \mathbb{E}[\Psi^{f}_{C_2}(C_1,C_2,m)] = 0.$$

The proofs can be found in Supplements B.10.9 and B.10.10

B.5.2 Dependent Mediators

We now assume the model :

$$\begin{aligned} Q &\sim \text{Uniform}(0,1) \\ T &\sim \text{Bernoulli}(0.5) \\ D|T,Q &\sim \text{Bernoulli}\left(\frac{4}{5} - \frac{3}{5}\frac{T+Q}{2}\right) \\ Y &= \alpha_Q Q + \alpha_D D + \alpha_T T + \alpha_{DT} DT + \alpha_{QT} QT + \varepsilon, \quad \mathbb{E}[\varepsilon|D,Q,T] = 0. \end{aligned}$$

We have

$$\begin{aligned} \Psi^{f^*}_{T\to Y|D,Q}(d,q) &= \alpha_T + \alpha_{DT} d + \alpha_{QT} q \\ \Psi^{f^*}_{T\to Y|Q}(q) &= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1,q] + \alpha_{QT} q + \alpha_D(\mathbb{E}[D|T=1,q] - \mathbb{E}[D|T=0,q]) \end{aligned}$$

$$
\begin{aligned}
&= \alpha_T + \alpha_{DT}(\frac{1}{2} - \frac{3q}{10}) + \alpha_{QT}q - \frac{3\alpha_D}{10} \\
\Psi^{f^*}_{T\to Y|D}(d) &= \alpha_T + \alpha_Q(\mathbb{E}[Q|T=1,d] - \mathbb{E}[Q|T=0,d]) + \alpha_{DT}d + \alpha_{QT}\mathbb{E}[Q|T=1,d] \\
&= \alpha_T - \frac{3\alpha_Q}{(13-6d)(7+6d)} + \alpha_{DT}d + \alpha_{QT}\frac{7-4d}{13-6d} \\
\Psi^{f^*}_{T\to Y|\emptyset} &= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT}q + \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{7\alpha_{DT}}{20} + \alpha_{QT}q - \frac{3\alpha_D}{10}
\end{aligned}
$$

where we used

$$
\begin{aligned}
\mathbb{E}[Q|T=1,d] &= \frac{7-4d}{13-6d} \\
\mathbb{E}[Q|T=0,d] &= \frac{4+2d}{7+6d} \\
\mathbb{E}[Q|T=1,d] - \mathbb{E}[Q|T=0,d] &= \frac{-3}{(13-6d)(7+6d)}
\end{aligned}
$$

which we prove from

$$
\begin{aligned}
p(d|q,t) &= dp(d=1|q,t) + (1-d)p(d=0|q,t) \\
&= d(\frac{4}{5} - \frac{3}{10}(t+q)) + (1-d)(\frac{1}{5} + \frac{3}{10}(t+q)) \\
p(d|t) &= \mathbb{E}[p(d|Q,t)|t] \\
&= \mathbb{E}[p(d|Q,t)] \text{ as } Q \perp\!\!\!\perp T \\
&= \frac{7}{20} + \frac{3t}{10} + \frac{3d}{5}(\frac{1}{2} - t) \\
p(q|d,t) &= \frac{p(d|q,t)p(q|t)p(t)}{p(d|t)p(t)} \\
&= \frac{p(d|q,t)}{p(d|t)} \\
&= \frac{1 + \frac{3}{2}(t+q) + 3d(1-t-q)}{\frac{7}{4} + \frac{3t}{2} + 3d(\frac{1}{2} - t)} \\
\mathbb{E}[Q|T=1,d] &= \int \frac{1 + \frac{3}{2} + \frac{3q}{2} - 3dq}{\frac{7}{4} + \frac{3}{2} - \frac{3d}{2}} q dq \\
&= \frac{\frac{1}{2}\frac{5}{2} + \frac{1}{3}(\frac{3}{2} - 3d)}{\frac{13}{4} - \frac{3d}{2}} \\
&= \frac{7-4d}{13-6d} \\
\mathbb{E}[Q|T=0,d] &= \int \frac{1 + 3d + q(\frac{3}{2} - 3d)}{\frac{7}{4} + \frac{3d}{2}} q dq
\end{aligned}
$$

$$= \frac{\frac{1}{2}(1+3d)+\frac{1}{3}(\frac{3}{2}-3d)}{\frac{7}{4}+\frac{3d}{2}}$$
$$= \frac{4+2d}{7+6d}$$

Thereby,

$$\begin{aligned}\Psi_Q^{f^*}(d,q) &= \Psi_{T\to Y|D,Q}^{f^*}(d,q) - \Psi_{T\to Y|D}^{f^*}(D) \\ &= \frac{3\alpha_Q}{(13-6d)(7+6d)} - \alpha_{QT}\frac{7-4d}{13-6d}\end{aligned}$$

Notably, we note that $\mathbb{E}[\Psi_Q^{f^*}(d,Q)] \neq 0$. Thereby, the local mediating effect as defined in Definition 5 is not able to isolate the absence of mediation from Q. However, defining

$$\Psi_{C_i,\text{alternative}}^{f}(q) := \Psi_{T\to Y|C_i}^{f}(c_i) - \Psi_{T\to Y|\emptyset}^{f} \tag{B.3}$$

we note that $\mathbb{E}[\Psi_{Q,\text{alternative}}^{f^*}(Q)] = 0$ and $\mathbb{E}[\Psi_{D,\text{alternative}}^{f^*}(D,q)] \neq 0$. Thereby, an alternative definition of the local mediating effect, as given in B.3, is able to isolate the absence of mediation from Q. This actually holds in a more general setting.

Property 8 (Ancestor of outcome). Let M be a post-treatment and pre-outcome variable. Assume that $M \perp\!\!\!\perp T$, or in other words M is not really a mediator. Then,

$$\mathbb{E}[\Psi_{T\to M\to Y}^{f}(M)] = 0$$

The proof can be found in Supplement B.10.11. Thereby, the original definition of the local mediating effect is not always suited for mediation analysis. However, picking up the alternative definition given above will mean we will not be able to use the same quantity for mediation, bias analysis and mediation analysis. Solving this dilemma is left for future work.

B.6 Generalisation to a Path of Length 3 or More

Assume the path p of length $L(p)+1$ is defined as $T \rightleftarrows C_{p(1)} \rightleftarrows C_{p(2)} \rightleftarrows \cdots \rightleftarrows C_{p(L(p))} \to Y$ where $\rightleftarrows$ is either $\leftarrow$ or $\to$ and there are no other paths from T into any of $C_{p(1)},\cdots,C_{p(L(p))}$ or from any of $C_{p(1)},\cdots,C_{p(L(p))}$ into Y. Then the path-wise Shapley effect with respect to p is

$$\Psi_p^f(c) = \Psi_{T\to Y|C_{S^*}}^{f}(c) - \Psi_{T\to Y|C_{S^*\setminus\{p(1),\ldots,p(L(p))\}}}^{f}(c_{S^*\setminus\{p(1),\ldots,p(L(p))\}}) \tag{B.4}$$

where the second term takes the coalition with all covariates in the path removed. Local confounding and moderating effects (Definitions 3 and 4, respectively) can be generalised

to paths p such that $T \leftarrow C_{p(1)}$, and local mediating effects (Definition 5) to paths p such that $T \rightarrow C_{p(1)}$. Lemma 1 and Property 5 can be generalised with these effects, by replacing covariates in the conditional independence statements with all covariates in either path.
If there are other paths of the form $T \rightarrow C_{p(k)}$ or $C_{p(k)} \rightarrow Y$, then the effect from Equation B.4 represents the effect of both p and the paths $T \rightleftarrows C_{p(1)} \rightleftarrows C_{p(2)} \rightleftarrows \cdots \rightleftarrows C_{p(k)} \rightarrow Y$ for all such k, as by grouping $C_{p(1)}, \cdots, C_{p(L(p))}$ into $C_{\{p(1),\dots,p(L(p))\}}$, all these paths would be merged into a single path $T \rightleftarrows C_{\{p(1),\dots,p(L(p))\}} \rightarrow Y$.

B.7 Further Experimental Results and Details

The code used for experiments is available at `https://github.com/oscarclivio/pwshap`.

B.7.1 Details of Experiments on Synthetic Datasets

Models : We model the outcome and propensity models using linear and logistic regression, respectively. Conditional distributions for PWSHAP are inferred by training an iterative imputer. Linear regressions with second order polynomial features are used to infer outcome models and logistic regressions for propensity models.
Causal Shapley : Regarding Causal Shapley, we model the
$\mathbb{E}\left[f\left(X_{\bar{S}}, x_{S\cup\{j\}}\right) \mid do\left(X_S = x_S\right)\right]$ term that is added and subtracted to obtain the direct and indirect effects (see Supplement B.3.3) by using an iterative imputer on the dataset obtained by removing the treatment column from the original dataset. do$-$distributions are modelled differently depending on the situation. They involve making variables independent from others. This is made by reshuffling the columns of these variables in the train set, then training the imputer on this modified dataset again.
Datasets : datasets are generated according to the models in Supplement B.4.2 for local bias analysis and in Supplement B.4.3 for local mediation analysis. 200 samples are generated and between training and testing sets as a 50/50 split. We change the links between treatment and covariates to model confounding and mediation, in the following way :

- Local bias analysis :
  - No confounders : $p(T=1|C_1,C_2) = 0.5$
  - C_1 is a confounder, C_2 is not : $p(T=1|C_1,C_2) = C_1$
  - C_1 and C_2 are confounders : $p(T=1|C_1,C_2) = C_1 C_2$
- Local mediation analysis :

Causal SHAP		PWSHAP			
ϕ_{direct}	ϕ_{indirect}	$\Psi^f_{\text{Occ}\to\text{Inc}}$	$\Psi^f_{\text{Occ}\leftarrow\text{Cntr}\to\text{ Inc}}$	$\Psi^f_{\substack{\text{Occ} \leftarrow \text{Cntr} \to \text{Capg} \to \text{Inc} \\ \text{Occ} \leftarrow \text{Cntr} \to \text{Inc}}}$	$\Psi^f_{\text{Relation:Occ}\to\text{Inc}}$
0.132 (0.22)	0 (0.027)	0.152 (1.98)	0.083	0.147 (1.13)	0.217 (0.831)

Table C2: Results for additional experiment on Census Income dataset where treatment is the individual's occupation. Note that $\Psi^f_{\substack{\text{Occ} \leftarrow \text{Cntr} \to \text{Capg} \to \text{Inc} \\ \text{Occ} \leftarrow \text{Cntr} \to \text{ Inc}}}$ is defined according to the generalization from Section B.6.

- No mediators : $Q|T \sim \text{Uniform}(0,1), D|T \sim \text{Binomial}(0.5)$.
- D is a mediator but not Q : $Q|T \sim \text{Uniform}(0,1), D|T \sim \text{Binomial}(\frac{4}{5} - \frac{3}{5}T)$
- Q and D are mediators : $D|T \sim \text{Binomial}(\frac{4}{5} - \frac{3}{5}T)$ and $Q = \frac{3}{5} \cdot T \cdot U + (1 - T) \cdot (\frac{3}{5} \cdot U + \frac{2}{5})$ where $U \sim \text{Uniform}(0,1)$

B.7.2 Further Results on the UCI Dataset

In an additional experiment, we consider the treatment of interest to the occupation of the individual. We focus on the question: *"Under what mechanisms did having a managerial occupation impact the model prediction?"*. Results comparing PWSHAP with Causal Shapley are shown in Table C2. Occupation seems to have a considerable impact throughout the cohort, with the base treatment effect showing that having a managerial job increases the predicted probability of high income by 15.2 points. Moderation by relationship status had a predominant effect in the model prediction for our individual. Being unmarried has increased the positive effect of having a managerial occupation by another 21.7 points. CausalSHAP does not capture this phenomenon as all the effect of occupation is deemed direct by definition. This further shows the high resolution of PWSHAP and impossibility of explaining. The local confounding effect of the pathway through country and capital gain also had an important impact.

B.7.3 Experimental Details on UCI

UCI dataset The UCI dataset –also known as "Census Income" dataset– predicts whether income exceeds \$50K/yr based on census data. It includes 11 features and 32,561 individuals. We select a random subsample of 5000 individuals and only consider the following 7 features: age, capital gain, native country, income, marital status, race and relationship. The occupation (Occ) and race variables were dichotomised, respectively into managerial/non-managerial and white/non-white. Other categorical features, namely native-country, marital status and relationship were encoded into numerical values. The URL for this dataset is `https://archive.ics.uci.edu/ml/datasets/adult`.

Pre-processing and model performance We use a Random Forest with 500 estimators and balanced class weights for all three models: the outcome model, the race propensity score model, the occupation propensity score model. We further used a Bayesian Ridge to learn the joint distribution on all covariates. Note that we could have inferred a propensity score model from the Bayesian Ridge but decided to fit a separate model so both weights and outcome models would be modelled with a classification tree.
The 5000 observations in the set were subsampled again 10 times, by taking out 20% of the sample. Each time, we use the subsample as both a training and testing set for our models. We further use it as a reference population. The goal is to explain how the model learned on this set, therefore using the same training set for testing isn't problematic. We ultimately want a model that has high accuracy whilst limiting the compute time. Means and standard deviations over all 10 subsamples are reported.
The accuracy of our outcome model is 0.827 and the AUC 0.944.
The accuracy of our race propensity score model is 0.894 and the AUC 0.950.
The accuracy of our occupation propensity score model is 0.757 and the AUC 0.863.

Computation and packages Random Forest and Bayesian Ridge were implemented using the `klearn package. We` ue the *dataset_fetcher* function from the Explanation GAME [71], available at:
`https://github.com/fiddler-labs/the-explanation-game-supplemental` (no license). Experiments were ran using a 2,6 GHz 6-Core Intel Core i7. The amount of compute time was approximately 2,500 clock-time seconds.

B.7.4 Experiments on the German Credit Dataset

Here, we apply PWSHAP to a local mediation analysis example on the German Credit Dataset [236]. This dataset assigns people described by age and gender (1 for male, 0 for female) and seeking a loan with a certain credit amount and a certain duration to a label about whether they are good customers (1 for yes, 0 for no). The DAG is given in Figure C1.
We consider a male candidate, aged 42, asking for a loan with credit amount 5507 euros and duration 24 years. This individual's loan application gets a probability of 0.713 from the black-box model, which is below the average probability of 0.728 in the training set. Results are shown in Figure C3. PWSHAP suggests that although the total effect of gender on decisions is slightly positive, the effect of gender on decisions through credit amount,

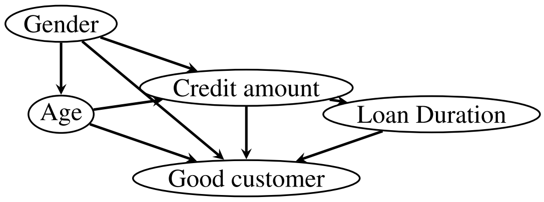


Figure C1: DAG of the German Credit Dataset.

Causal SHAP	ϕ_{direct}	0.025
	ϕ_{indirect}	0.004
PWSHAP	$\Psi_{\text{Gender}\xrightarrow{\text{total}}\text{Good}}$	0.069
	$\Psi_{\text{Gender}\to\text{Amount}\to\text{Good}}$	-0.1819
	$\Psi_{\text{Gender}\to\text{Amount}\to\text{Duration}\to\text{Good}} + \Psi_{\text{Gender}\to\text{Amount}\to\text{Good}}$	-0.2371

Table C3: Results on the German Credit Dataset. Note that the last PWSHAP effect follows the generalization from Appendix B.6.

or through credit amount and loan duration, is negative. This suggests that although being male might give a slight advantage to this individual, it is compensated by high credit amount and loan duration, which are in the 65-th and 81-th percentiles, respectively, producing a relatively bad prediction compared to the average. Causal Shapley notes both a positive direct effect and a small positive indirect effect of gender on credit amount, which is not consistent with the finding that the predicted probability of being a good customer is lower than average for this individual.

B.7.5 Computational Complexity

Complexity of PWSHAP effects : For ease of notation, assume that the size of the data and the number of Monte-Carlo samples for expectations are all of size $\mathcal{O}(n)$. We denote by p the number of covariates. We also assume the black-box model and the propensity score model require $\mathcal{O}(p)$ steps to be evaluated, as e.g. in linear/logistic regression, that we have access to all conditional distributions and that sampling any feature from other features can be done in $\mathcal{O}(p)$ too (e.g. if, again, all conditional distributions are GLMs). For any sample i and any path of length $l+1$ (with the treatment, l covariates and the outcome), computing the path-wise PWSHAP effect mostly requires evaluating expectations of the black-box or propensity score model, where :

- we take $\mathcal{O}(n)$ Monte-Carlo samples of l (when the treatment is observed) or $l+1$ (when the treatment is missing) variables considered to be missing from other covariates (complexity $\mathcal{O}(nlp)$) ;

- we evaluate and aggregate the black-box for all these imputed points (complexity $\mathcal{O}(np)$).

Thus, the PWSHAP effect on a path of size $l+1$ for one sample has a complexity $\mathcal{O}(nlp)$. Aggregated over all samples, the PWSHAP of this path is of complexity $\mathcal{O}(n^2lp)$. Aggregated over all paths for all l, we have a complexity of at most $\mathcal{O}(n^2p^22^p)$. However, this can be reduced to $\mathcal{O}(n^2p^2)$ if the number of paths is small.
All of this assumes that we have access to conditional distributions. In practice, we do not and resort to an iterative imputer that relies on the MICE algorithm. Although, we could not find details on its complexity, scikit-learn documentation[2] suggests that imputation and thus sampling from the conditional distribution can be done in polynomial factors w.r.t. n and p too. If training the imputer is polynomial, then the overall complexity of computing all PWSHAP effects will remain polynomial in n and p.
Thus, the method might not scale well with p, but with few paths in the DAG it would scale better than exact classical observational/marginal Shapley values whose complexities have a 2^p factor.
Complexity of other Shapley value methods : with the same assumptions, exact on-manifold and off-manifold Shapley values have a $\mathcal{O}(n^2p^22^p)$ or $\mathcal{O}(n^2p2^p)$ complexity, respectively. Although the general computational complexity of the approximation made by KernelSHAP [57] is not explicited in their work, we expect it to give better complexity at the cost of further approximating Shapley values compared to direct evaluation of expectations. TreeSHAP (an explanation model for tree ensemble models only) [237] reduces the 2^p factor by a square factor in the maximal depth of trees, however this method isn't model agnostic like most Shapley approaches, including PWSHAP.

B.8 Running Example

Using the causal structure described in the DAG in Figure C2 as a running example, we illustrate our concepts. Here, we consider three variables: *Sex* (binary, denoting the female sex), *Surg* (binary, denoting the execution of a surgery) and *Preg* (binary, denoting whether the subject is pregnant). We denote *Death* the logit of the probability of death as a result of the surgery, of its absence. We analyse the effect of the sensitive attribute *Sex* on the *Death* outcome, with regards to model fairness. The model is as follows:
We outline a few definitions and concepts applied to this example:

[2]`https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html`

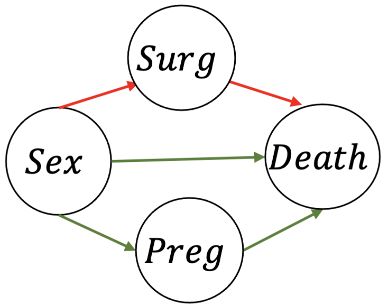


Figure C2: DAG for the running example

- On-manifold Shapley value for Sex on the empty coalition:

$$\phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) = (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{sex}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{sex}]) \\ - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}])$$

- Off-manifold Shapley value for Sex on the empty coalition:

$$\begin{aligned}\phi^{f,\text{off-manifold}}_{\text{Sex},\emptyset}(\text{sex}) &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= \alpha_{\text{Sex}}(\text{Sex} - \mathbb{E}[\text{Sex}])\end{aligned}$$

- Causal Shapley value for Sex on the empty coalition:

$$\begin{aligned}\phi^{f,\text{causal}}_{\text{Sex},\emptyset}(\text{sex}) &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{do}(\text{Sex})] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{do}(\text{Sex})]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{sex}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{sex}]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= \phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) \quad \text{for this value in this specific example.}\end{aligned}$$

- Coalition-specific Shapley effect (for Sex, our treatment of interest) on the empty coalition: one can note that the on-manifold Shapley value can be decomposed as

$$\phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) = (\text{sex} - \mathbb{E}[\text{Sex}])(\alpha_{\text{Sex}} + \alpha_{\text{Preg}} \cdot p_{\text{Preg}} + \alpha_{\text{Surg}} \cdot p_{\text{Surg}})$$

  where the second factor on the right-hand side is the coalition-specific Shapley effect, which here is written as

$$\mathit{effect}^{f}_{\text{Sex}\to\text{Death}|\emptyset} = \mathbb{E}[f(\text{sex}=1, \text{Surg}, \text{Preg})] - \mathbb{E}[f(\text{sex}=0, \text{Surg}, \text{Preg})]$$

- Path-wise Shapley effect of Surg: here it is expressed as

$$\mathit{effect}^{f}_{\text{Surg}} = \mathit{effect}^{f}_{\text{Sex}\to\text{Death}|\text{Surg},\text{Preg}}(\text{surg}, \text{preg}) - \mathit{effect}^{f}_{\text{Sex}\to\text{Death}|\text{Preg}}(\text{preg})$$

  We can interpret this as an effect through the path Sex $\to$ Surg $\to$ Death, re-noting it as $\mathit{effect}^{f}_{\text{Sex}\to\text{Surg}\to\text{Death}}$.

B.9 Experiments on Synthetic Datasets

We now illustrate PWSHAP's ability to capture confounding and mediation on a set of synthetic datasets, some of which being drawn from the examples developed above. We infer path-specific Shapley effects following the procedure described in 4.3.4 and compute their averages $\bar{\Psi}$ across the testing set. To mitigate variations in results due to the scale of the outcome, we present these averages divided by the empirical standard deviation of outcome σ_Y, computed across the training set. Conditional distributions are inferred by training an iterative imputer, linear regressions with 2nd order polynomial features are used to infer outcome models and logistic regressions for propensity models, all of this corresponding in part to a typical user behaviour, agnostic to the examples outlined below.
Local bias analysis, single pre-treatment covariate : we assume that the logit of the probability of death ("Death") is known and depends linearly on an age variable ("Age") and a surgery indicator ("Surg") which is our treatment of interest, in a linear model as $\text{Death} = \alpha_{\text{Age}}\text{Age} + \alpha_{\text{Surg}}\text{Surg}$.
We look at scenarios where the surgery indicator is randomised ("Rand. surg."), or it depends on the age variable ("Non-rand. surg."), and where the outcome model is the true model ("True model"), depending on both age and surgery, or is made dependent on the surgery variable only, becoming unconfounded ("Unconfounded model"). We evaluate the local confounding effect of the age variable. Results are presented in Table C4. The pathwise-effect-to-standard-deviation ratio is significantly lower in absolute value for the two scenarios where surgery is not confounded by age than for the scenario where it is. These results are consistent with Lemma 1. We compute analoguous ratios for both Causal Shapley and observational Shapley values. We notably observe that for any scenario, these ratios are always near-zero. This shows that Causal Shapley and observational Shapley are unsuited to distinguish confounding from its absence in this example with a linear model without interactions. More specifically, for Causal Shapley, this somewhat surprising finding stems from the Shapley values being identical for both a randomised and a non-randomised treatment, as a result of the do-operations. For both CS and SHAP, this finding is further caused by the linear model without no interaction. This can be generalised to any number of independent pre-treatment covariates in the absence of interactions.
Local bias analysis, two pre-treatment covariates: We now assume a model with two pre-treatment covariates C_1 and C_2 as described in Section 4.6.2. We examine the local confounding effects of C_1 and C_2.
Using the functional form of the propensity score, we consider three scenarios: (1) neither C_1 nor C_2 are confounders, i.e., $\mathbb{P}(T = 1 \mid C_1, C_2) = 0.5$; (2) C_1 is a confounder but C_2 is not, i.e., $\mathbb{P}(T = 1 \mid C_1, C_2) = C_1$; and (3) both C_1 and C_2 are confounders, i.e., $\mathbb{P}(T = 1 \mid C_1, C_2) = C_1 C_2$.

	Rand. surg., True model	Non-rand. surg., true model	Non-rand. surg., unconfounded model
$\bar{\psi}^{f}_{\text{Age}\leftarrow\text{Surg}\rightarrow\text{Death}}/\sigma_Y$	0.005	0.348	-0.044
$\bar{\phi}^{f,\text{CS}}_{\text{Surg}}/\sigma_Y$	0.026	0.027	-0.009
$\bar{\phi}^{f,\text{SHAP}}_{\text{Surg}}/\sigma_Y$	-0.028	0.035	0.007

Table C4: Results on local bias analysis, one single pre-treatment covariate.

Scenario	C_1, C_2 non-confounders	C_1 confounder, C_2 non-confounder	C_1, C_2 confounders
$\bar{\Psi}^{f}_{T\leftarrow C_1\rightarrow Y}/\sigma_Y$	-0.024	-0.589	-0.387
$\bar{\Psi}^{f}_{T\leftarrow C_2\rightarrow Y}/\sigma_Y$	-0.051	-0.036	-0.095

Table C5: Results on local bias analysis, two synthetic pre-treatment covariates.

Results are presented in Table C5. We note that pathwise-effect-to-standard-deviation ratios are significantly higher in absolute value for variables that are confounders compared to those that are not. An exception is observed for C_2 in the third scenario, where its ratio remains higher than in scenarios where C_2 is not a confounder.

These findings further support the validity of PWSHAP's local confounding effect in isolating individual non-confounders in pre-treatment covariates, in accordance with Lemma 1.

Local mediation analysis, two post-treatment covariates : we now assume the model for college admissions described in Section B.4.3 where a gender indicator T can influence exam results Q and the choice of a competitive department D, and all of T, Q, D determine the outcome Y taken as the logit of the probability of admission. We look at the local mediating effects of Q and D, in three scenarios : (i) T influences neither Q or D, so neither is a mediator, (ii) T influences D but not Q, so the former is a mediator but the latter is not, (iii) T influences both Q and D, so both are mediators. Results are presented in Table C6. We note that pathwise-effect-to-standard-deviation ratios are significantly higher in absolute values for variables that are mediators than variables that are not mediators. All of this further supports the validity of PWSHAP's local mediating effect to isolate individual non-mediators in pre-treatment covariates, in accordance with Property 5.

Scenario	Q, D non-mediators	Q mediator, D non-mediator	Q, D mediators
$\bar{\Psi}^{f}_{T\rightarrow Q\rightarrow Y}/\sigma_Y$	-0.016	0.015	2.074
$\bar{\Psi}^{f}_{T\rightarrow D\rightarrow Y}/\sigma_Y$	-0.022	0.577	0.953

Table C6: Results on local mediation analysis, two synthetic post-treatment covariates.

Empirically, all of these results suggest that having a ratio $\bar{\Psi}^f_{C_i}/\sigma_Y$ below 0.05 in absolute value indicates that some given statistical dependencies of interest like confounding or mediation might be cut off. We do not have a clear recommendation for such a threshold however, especially as these ratios or simply the path-specific effects do not reach negligible values, eg below 10^{-5}, even in the absence of confounding or mediation. This can be understood as a limitation of our approach. We do, however, advocate the use of some normalisation, eg dividing by the standard deviation of outcomes σ_Y, to isolate near-zero pathwise effects indicating the absence of confounding or mediation, as these pathwise effects can otherwise be made arbitrarily large or small by scaling the outcomes. Other than the outcome, the propensity score model can also lead to less stable estimates of pathwise Shapley effects when its values are closer to 0 or when it is not correctly fitted. On the other hand, PWSHAP values are estimated in a manner agnostic to the example developed, while Causal Shapley must be computed in a different way depending on the DAG under scrutiny. Further, we have found an example where Causal Shapley is not suited to provide causal interpretations of the data under scrutiny, whereas an agnostic approach for PWSHAP yields causal interpretations that are consistent with theoretical developments in all above examples.

B.10 Proofs of Properties and Lemmas

B.10.1 Proof of Property 1

It suffices to show that

$$\phi^f_{T,S}(c,t) = w^*_S \times (v_f(S \cup \{T\}, c, 1) - v_f(S \cup \{T\}, c, 0))$$

where

$$\begin{aligned} v_f(S \cup \{T\}, c, t) &= v_f(S \cup \{T\}, c_S, t) = \mathbb{E}_{p(C_{\bar{S}}|c_S,t)}[f(c_S, C_{\bar{S}}, t)] \\ v_f(S, c) &= \mathbb{E}_{p(C_{\bar{S}},T|c_S)}[f(c_S, C_{\bar{S}}, T)] \\ w^*_S = w^*(c_S, t) &= t - P(T = 1|C_S = c_S) \end{aligned}$$

In other words we want to show that

$$\begin{aligned} \phi^f_{T,S}(c,t) = (t - p(T = 1|C_S = c_S)) \times (&\mathbb{E}_{p(C_{\bar{S}}|c_S,1)}[f(c_S, C_{\bar{S}}, 1)] \\ &- \mathbb{E}_{p(C_{\bar{S}}|c_S,0)}[f(c_S, C_{\bar{S}}, 0)]) \end{aligned}$$

We note that

$$v_f(S,c) = \mathbb{E}[f(c_S, C_{\bar{S}}, T)|c_S = c_S]$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}[f(c_S, C_{\bar{S}}, T)|c_S = c_S, T]|c_S = c_S] \\
&= p(T = t|c_S) \times \mathbb{E}[f(c_S, C_{\bar{S}}, t)|c_S = c_S, T = t] \\
&\quad + p(T = 1-t|c_S) \times \mathbb{E}[f(c_S, C_{\bar{S}}, 1-t)|c_S = c_S, T = 1-t] \\
&= p(T = t|c_S) \times v_f(S \cup \{T\}, c, t) \\
&\quad + p(T = 1-t|c_S) \times v_f(S \cup \{T\}, c, 1-t)
\end{aligned}
$$

and, from $1 = p(T = t|c_S) + p(T = 1-t|c_S)$,

$$
\begin{aligned}
&v_f(S \cup \{T\}, c, t) - v_f(S, c) \\
&= p(T = t|c_S) v_f(S \cup \{T\}, c, t) + p(T = 1-t|c_S) v_f(S \cup \{T\}, c, t) \\
&\quad - p(T = t|c_S) v_f(S \cup \{T\}, c, t) - p(T = 1-t|c_S) v_f(S \cup \{T\}, c, 1-t).
\end{aligned}
$$

So terms in $p(T = t|c_S)$ cancel out and we get:

$$
\begin{aligned}
&v_f(S \cup \{T\}, c, t) - v_f(S, c) \\
&\quad = p(T = 1-t|c_S) \times [v_f(S \cup \{T\}, c, t) - v_f(S \cup \{T\}, c, 1-t)] \\
&\quad = (t - \pi_S^*(c_S)) \times [v_f(S \cup \{T\}, c, 1) - v_f(S \cup \{T\}, c, 0)]
\end{aligned}
$$

where $\pi_S^*(c_S) = P(T = 1|C_s = c_S)$ and the second equality comes from $t \in \{0, 1\}$.

B.10.2 Proof of Property 2

Let c_{-i} be a value of C_{-i}. It suffices to show that $\mathbb{E}[\Psi^f_{T \to Y|C_i, C_{-i}}(C_i, c_{-i})|C_{-i} = c_{-i}] = \Psi^f_{T \to Y|C_{-i}}(c_{-i})$, which is true as

$$
\begin{aligned}
&\mathbb{E}[\Psi^f_{T \to Y|C_i, C_{-i}}(C_i, c_{-i})|C_{-i} = c_{-i}] \\
&= \int_{c_i} (v_f(S^* \cup \{T\}, c_i, c_{-i}, 1) - v_f(S^* \cup \{T\}, c_i, c_{-i}, 0)) \mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} (f(c_i, c_{-i}, 1) - f(c_i, c_{-i}, 0)) \mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} f(c_i, c_{-i}, 1) \mathrm{d}p(c_i|c_{-i}) - \int_{c_i} f(c_i, c_{-i}, 0) \mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} f(c_i, c_{-i}, 1) \mathrm{d}p(c_i|c_{-i}, t = 1) - \int_{c_i} f(c_i, c_{-i}, 0) \mathrm{d}p(c_i|c_{-i}, t = 0) \text{ as } T \perp\!\!\!\perp C_i | C_{-i} \\
&= \mathbb{E}[f(C_i, c_{-i}, 1)|C_{-i} = c_{-i}, T = 1] - \mathbb{E}[f(C_i, c_{-i}, 0)|C_{-i} = c_{-i}, T = 0] \\
&= v_f((S^* \backslash \{i\}) \cup \{T\}, c_{-i}, 1) - v_f((S^* \backslash \{i\}) \cup \{T\}, c_{-i}, 0) \\
&= \Psi^f_{T \to Y|C_{-i}}(c_{-i})
\end{aligned}
$$

which completes the proof.

B.10.3 Proof of Property 3

We assume that

$$\forall c,t,N, |\hat{f}_N(c,t) - f^*(c,t)| \leq e_N^{\text{outcome}}$$

Then, for any coalition S without T, c and N,

$$\begin{aligned}
&|\Psi^{\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c)| \\
&\leq |(\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S,C_{\bar{S}},1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S,C_{\bar{S}},0)]) \\
&\quad - (\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[f^*(c_S,C_{\bar{S}},1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[f^*(c_S,C_{\bar{S}},0)])| \\
&= |(\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S,C_{\bar{S}},1) - f^*(c_S,C_{\bar{S}},1)] \\
&\quad\; - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S,C_{\bar{S}},0) - f^*(c_S,C_{\bar{S}},0)])| \\
&\leq |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S,C_{\bar{S}},1) - f^*(c_S,C_{\bar{S}},1)]| \\
&\quad\; + |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S,C_{\bar{S}},0) - f^*(c_S,C_{\bar{S}},0)]| \text{ from the triangle inequality} \\
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[|\hat{f}_N(c_S,C_{\bar{S}},1) - f^*(c_S,C_{\bar{S}},1)|] \\
&\quad\; + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[|\hat{f}_N(c_S,C_{\bar{S}},0) - f^*(c_S,C_{\bar{S}},0)|] \text{ from Jensen's inequality} \\
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[e_N^{\text{outcome}}] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[e_N^{\text{outcome}}] \text{ by assumption} \\
&= 2e_N^{\text{outcome}}
\end{aligned}$$

which shows bound 1. Now, let S as before, c, t, N,

$$\begin{aligned}
&|\phi^{\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)| \\
&= |w^*_S(t,c_S) \cdot (\Psi^{\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c))| \\
&= |w^*_S(t,c_S)||\Psi^{\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c)| \\
&\leq |\Psi^{\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c)| \text{ as } |w^*_S(t,c_S)| \leq 1 \\
&\leq 2e_N^{\text{outcome}} \text{ from the previous bound}
\end{aligned}$$

which proves bound 2. Now, for any covariate feature i as before, c,N

$$\begin{aligned}
&|\Psi^{\hat{f}_N}_{C_i}(c) - \Psi^{f^*}_{C_i}(c)| \\
&= |\Psi^{\hat{f}_N}_{T\to Y|C_{S^*}}(c) - \Psi^{\hat{f}_N}_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}}) - (\Psi^{f^*}_{T\to Y|C_{S^*}}(c) - \Psi^{f^*}_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}}))| \\
&\leq |\Psi^{\hat{f}_N}_{T\to Y|C_{S^*}}(c) - \Psi^{f^*}_{T\to Y|C_{S^*}}(c)| \\
&\quad + |\Psi^{\hat{f}_N}_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}}) - \Psi^{f^*}_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}})|
\end{aligned}$$

from the triangle inequality

$\leq 2e_N^{\text{outcome}} + 2e_N^{\text{outcome}}$ from the bound on the coalition-wise Shapley effect

which proves bound 3.

B.10.4 Proof of Property 4

We assume that

$$\forall S \text{ s.t. } T \notin S, c, N, |\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)| \leq e_N^{\text{Shap}},$$

that the arbitrary propensity score model π^N and π^* verify ε-strong overlap, i.e. $\varepsilon \leq \pi^N \leq 1-\varepsilon$, and $\varepsilon \leq \pi^* \leq 1-\varepsilon$, and that we have $\forall c, N, |\pi^N(c) - \pi^*(c)| \leq e_N^{\text{propensity}}$.
Then, for any S s.t. $T \notin S$, we note that ε-strong overlap for π^N and π^* implies ε-strong overlap for $\pi_S^N(c_S) := \mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^N(c_S, C_{\bar{S}})]$ and $P(T=1|C_S=c_S) = \mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^*(c_S, C_{\bar{S}})]$ (by taking the expectation w.r.t. $p(C_{\bar{S}}|c_S)$) and also Jensen's inequality yields

$$\forall c, N, |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^N(c_S, C_{\bar{S}})] - P(T=1|C_S=c_S)| \leq e_N^{\text{propensity}},$$

which further gives $\forall t, c, N, |w_S^N(c,t) - w_S^*(c,t)| \leq e_N^{\text{propensity}}$.
So for any S s.t. $T \notin S$, c, N,

$$\begin{aligned}|w_S^N(c,0)| &= |-\pi_S^N(c)| = \pi_S^N(c) \geq \varepsilon \\ |w_S^N(c,1)| &= |1-\pi_S^N(c)| = 1-\pi_S^N(c) \geq \varepsilon\end{aligned}$$

Thereby, for any S s.t. $T \notin S$, t, c, N,

$$\frac{1}{|w_S^N(c,t)|} \leq \frac{1}{\varepsilon}$$

and, similarly,

$$\frac{1}{|w_S^*(c,t)|} \leq \frac{1}{\varepsilon}.$$

We also show that $|\phi_{T,S}^{f^*}(c,t)| \leq 2||f^*||_\infty$: indeed,

$$\begin{aligned}|\phi_{T,S}^{f^*}(c,t)| &= |w_S^*(c,t)||\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[f^*(c_S, C_{\bar{S}}, 1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[f^*(c_S, C_{\bar{S}}, 0)]| \\ &\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[|f^*(c_S, C_{\bar{S}}, 1)|] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[|f^*(c_S, C_{\bar{S}}, 0)|] \\ &\qquad \text{from } |w_S^*(c,t)| \leq 1 \text{ and the triangle inequality and Jensen's inequality} \\ &\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[||f^*||_\infty] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[||f^*||_\infty] \\ &= 2||f^*||_\infty\end{aligned}$$

In the end, we have

$$
\begin{aligned}
&|\hat{\Psi}^{N,\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c)| \\
&= |\frac{\hat{\phi}^{N,\hat{f}_N}_{T,S}(c,t)}{w^N_S(c,t)} - \frac{\phi^{f^*}_{T,S}(c,t)}{w^*_S(c,t)}| \\
&= |\frac{\hat{\phi}^{N,\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)}{w^N_S(c,t)} + \phi^{f^*}_{T,S}(c,t)(\frac{1}{w^N_S(c,t)} - \frac{1}{w^*_S(c,t)})| \\
&= |\frac{\hat{\phi}^{N,\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)}{w^N_S(c,t)} + \phi^{f^*}_{T,S}(c,t)\frac{w^*_S(c,t) - w^N_S(c,t)}{w^*_S(c,t)w^N_S(c,t)}| \\
&\leq \frac{|\hat{\phi}^{N,\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)|}{|w^N_S(c,t)|} + |\phi^{f^*}_{T,S}(c,t)|\frac{|w^*_S(c,t) - w^N_S(c,t)|}{|w^*_S(c,t)||w^N_S(c,t)|} \\
&\text{from Jensen's inequality} \\
&\leq \frac{2e^{\text{Shapley}_N}}{\varepsilon} + 2||f^*||_\infty \cdot \frac{e^{\text{propensity}}_N}{\varepsilon^2}
\end{aligned}
$$

which proves bound 4. Bound 5 is proven similarly to bound 3 above.

B.10.5 Proof of Lemma 1.

If unconfoundess w.r.t. C_1, C_2 holds then

$$\mathbb{E}[\Psi^{f^*}_{T\to Y|C_1,C_2}(C_1,C_2)] = \mathbb{E}[\mathbb{E}[Y|T=1,C_1,C_2] - \mathbb{E}[Y|T=0,C_1,C_2]] = \text{ATE}.$$

If unconfoundess w.r.t. C_1 also holds then

$$\mathbb{E}[\Psi^{f^*}_{T\to Y|C_1}(C_1)] = \mathbb{E}[\mathbb{E}[Y|T=1,C_1] - \mathbb{E}[Y|T=0,C_1]] = \text{ATE}$$

In the end,

$$
\begin{aligned}
\mathbb{E}[\Psi^{f^*}_{T\leftarrow C_2\to Y}(C_1,C_2)] = \mathbb{E}[\Psi^{f^*}_{T\to Y|C_1,C_2}(C_1,C_2) - \Psi^{f^*}_{T\to Y|C_1}(C_1)] \\
= \text{ATE} - \text{ATE} = 0.
\end{aligned}
$$

We also show that

- $\forall t, Y(t) \perp\!\!\!\perp T \mid C_1, C_2$
- $T \perp\!\!\!\perp C_2 \mid C_1$

imply $\forall t, Y(t) \perp\!\!\!\perp T \mid C_1$. Indeed, let $t \in \{0,1\}$,

$$P(Y(t),T|C_1) = \int P(Y(t),T|C_1,C_2=c_2)dP_{C_2|C_1}(c_2)$$

$$
\begin{aligned}
&= \int P(Y(t)|T, C_1, C_2 = c_2) P(T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&= \int P(Y(t)|C_1, C_2 = c_2) P(T|C_1) dP_{C_2|C_1}(c_2) \text{ due to assumptions} \\
&= P(T|C_1) \cdot \int P(Y(t)|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&= P(T|C_1) P(Y(t)|C_1)
\end{aligned}
$$

which completes this part of the proof. Finally, we show that

- $\forall t, Y(t) \perp\!\!\!\perp T \mid C_1, C_2$
- $\forall t, Y(t) \perp\!\!\!\perp C_2 \mid C_1$

imply $\forall t, Y(t) \perp\!\!\!\perp T \mid C_1$. Indeed, let $t \in \{0, 1\}$,

$$
\begin{aligned}
P(Y(t), T|C_1) &= \int P(Y(t), T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&= \int P(Y(t)|T, C_1, C_2 = c_2) P(T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&= \int P(Y(t)|C_1, C_2 = c_2) P(T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&\quad \text{due to the first assumption} \\
&= \int P(Y(t)|C_1) P(T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&\quad \text{due to the second assumption} \\
&= P(Y(t)|C_1) \cdot \int P(T|C_1, C_2 = c_2) dP_{C_2|C_1}(c_2) \\
&= P(Y(t)|C_1) P(T|C_1)
\end{aligned}
$$

which completes the proof.
Remark : the condition $Y \perp\!\!\!\perp C_2|C_1, T$ (or the outcome model not depending on C_2) is *not* a sufficient condition for unconfoundness w.r.t. C_1 to hold if unconfoundness w.r.t. C_1, C_2 holds. Indeed, take C_2 to be binary, T to be the exact copy of C_2, and $\forall t, Y(t) = C_1 + t \cdot C_2$. Then it turns out that $Y = C_1$ so $Y \perp\!\!\!\perp C_2|C_1, T$ holds. However, unconfoundness w.r.t. C_1, C_2 holds but not unconfoundness w.r.t. C_1 alone. We note however that the overlap assumption is violated in this example. If all assumptions making identification of $P(Y(t)|C_1, C_2) = P(Y|T = t, C_1, C_2)$ are possible, then the condition $Y \perp\!\!\!\perp C_2|C_1, T$ and the above condition $\forall t, Y(t) \perp\!\!\!\perp C_2 \mid C_1$ are equivalent.

B.10.6 Proof of Property 5

Let c and m_1 be values of C and M_1, respectively. We note that

$$\begin{aligned}
&\mathbb{E}[\Psi^f_{T\to M_2\to Y}(c,m_1,M_2)|C=c] \\
&=\mathbb{E}[\Psi^f_{T\to M_2\to Y}(c,m_1,M_2)|C=c,M_1=m_1] \text{ as } M_1 \perp\!\!\!\perp M_2|C \\
&=0 \qquad \text{from Property 2 as } M_1 \perp\!\!\!\perp T|M_2,C
\end{aligned}$$

which completes the proof.

B.10.7 Proof of Property 6

If a latent variable generates all pre-treatment covariates of T, then we can factorise the distribution of $(T,C_S,C_{\bar{S}},f(C_{\bar{S}},c_S,t))$ in the ADMG with those variables as nodes and edges $C_S \leftrightarrow C_{\bar{S}} \to f(C_{\bar{S}},c_S,t)$ and $C_{\bar{S}} \to T \leftarrow C_S$. We aim to apply rule 3 of Pearl's do-calculus. If we remove edges pointing into C_S and T, we obtain an ADMG with only the edge $C_{\bar{S}} \to f(C_{\bar{S}},c_S,t)$. In this graph T and $f(C_{\bar{S}},c_S,t)$ are m-separated by $C_{\bar{S}}$. Therefore, rule of 3 of do-calculus applies and we can remove the $T=t$ term in the do-operator of the left-hand term, yielding the right-hand term. Hence the indirect effect is zero.

B.10.8 Proof of Corollary 1

We note that

$$\begin{aligned}
\mathbb{E}[\Psi^f_{T\leftarrow C_2\to Y}(C_1,C_2)] &= \mathbb{E}[\mathbb{E}[\Psi^f_{T\leftarrow C_2\to Y}(C_1,C_2)|C_1]] \text{ from the tower property} \\
&= \mathbb{E}[0] \text{ from Property 2 as } C_2 \perp\!\!\!\perp T|C_1 \\
&= 0
\end{aligned}$$

B.10.9 Proof of Lemma 3.

Let m be a value of M. If $\mathscr{H}(C)$ holds then for any $t=0,1$

$$\begin{aligned}
\mathbb{E}[Y(t,m)] &= \mathbb{E}[\mathbb{E}[Y(t,m)|C]] \\
&= \mathbb{E}[\mathbb{E}[Y(t,m)|C,t]] \text{ from } Y(t,m) \perp\!\!\!\perp T|C \\
&= \mathbb{E}[\mathbb{E}[Y(t,m)|C,t,m]] \text{ from } Y(t,m) \perp\!\!\!\perp M|T,C \\
&= \mathbb{E}[\mathbb{E}[Y|C,t,m]] \text{ from consistency.}
\end{aligned}$$

so

$$\begin{aligned}\mathbb{E}[\Psi^{f^*}_{T\to Y|C,M}(C,m)] &= \mathbb{E}[\mathbb{E}[Y|C,t=1,m]]-\mathbb{E}[\mathbb{E}[Y|C,t=0,m]]\\ &= \mathbb{E}[Y(1,m)]-\mathbb{E}[Y(0,m)] \text{ from the above}\\ &= \mathrm{CDE}(m).\end{aligned}$$

So if both $\mathscr{H}(C_1,C_2)$ and $\mathscr{H}(C_1)$ hold then

$$\begin{aligned}\mathbb{E}[\Psi^{f^*}_{C_2}(C_1,C_2,m)] &= \mathbb{E}[\Psi^{f^*}_{T\to Y|C_1,C_2,M}(C_1,C_2,m)-\Psi^{f^*}_{T\to Y|C_1,M}(C_1,m)]\\ &= \mathrm{CDE}(m)-\mathrm{CDE}(m)\\ &= 0.\end{aligned}$$

B.10.10 Proof of Corollary 2

First, let's note that $C_2 \perp\!\!\!\perp T,M|C_1$ implies $C_2 \perp\!\!\!\perp M|C_1$ and $C_2 \perp\!\!\!\perp T|C_1,M$. Let m be a value of M. We note that

$$\begin{aligned}\mathbb{E}[\Psi^{f}_{C_2}(C_1,C_2)] &= \mathbb{E}[\mathbb{E}[\Psi^{f}_{C_2}(C_1,C_2)|C_1]] \text{ from the tower property}\\ &= \mathbb{E}[\mathbb{E}[\Psi^{f}_{C_2}(C_1,C_2)|C_1,M]] \text{ from the property } C_2 \perp\!\!\!\perp M|C_1\\ &= \mathbb{E}[0] \text{ from Property 2 as } C_2 \perp\!\!\!\perp T|C_1,M\\ &= 0\end{aligned}$$

which completes the proof.

B.10.11 Proof of Property 8

It suffices to show that $\mathbb{E}[\Psi^{f}_{T\to Y|M}(M)] = \Psi^{f}_{T\to Y|\emptyset}$, which is true as

$$\begin{aligned}&\mathbb{E}[\Psi^{f}_{T\to Y|M}(M)]\\ &= \int_m (v_f(\{M,T\},m,t=1)-v_f(\{M,T\},m,t=0))\mathrm{d}p(m)\\ &= \int_m (\mathbb{E}[f(C,m,t=1)|m,t=1]-\mathbb{E}[f(C,m,t=0)|m,t=0)\mathrm{d}p(m)\\ &= \int_m \mathbb{E}[f(C,m,t=1)|m,t=1]\mathrm{d}p(m)-\int_m \mathbb{E}[f(C,m,t=0)|m,t=0]\mathrm{d}p(m)\\ &= \int_m \mathbb{E}[f(C,m,t=1)|m,t=1]\mathrm{d}p(m|t=1)-\mathbb{E}[f(C,m,t=1)|m,t=0]\mathrm{d}p(m|t=0)\\ &\text{as } M \perp\!\!\!\perp T\\ &= \mathbb{E}[f(C,M,t=1)|t=1]-\mathbb{E}[f(C,M,t=1)|t=0]\end{aligned}$$

$$
\begin{aligned}
&= v_f(T, t=1) - v_f(T, t=0) \\
&= \Psi^f_{T \to Y|\emptyset}
\end{aligned}
$$

which completes the proof.

Appendix C

Supplementary Materials: Hierarchical Bias-Driven Stratification for Interpretable Causal Effect Estimation

C.1 Causal inference

Confounder A confounder is a variable that is associated with both the treatment and the outcome, causing a spurious correlation. For instance, summer is associated with eating ice cream and getting sunburns, but there is no causal relationship between the two.

Propensity score model A propensity score model is a function that predicts treatment from the observed covariates i.e. $P(T=1|C=c)$ for a binary treatment T and a covariate vector C.

Potential outcome As defined by the Rubin causal model [109], a potential outcome $Y(t)$ is the value that Y would take if T were set by (hypothetical) intervention to the value t.

Identification assumptions Inference is possible under three identification assumptions.

- **No interference** For a given individual i, this assumption implies that $Y_i(t)$ represents the value that Y would have taken for individual i if T had been set to t for individual i, i.e the potential value of Y_i if T_i had been set to t.
- **Consistency** For a given individual i, $T_i = t \Rightarrow Y_i = Y_i(t)$. This means that for individuals who actually received treatment level t, their observed outcome is the same as what it would have been had they received treatment level t via an hypothetical intervention.

- **Conditional exchangeability** For a given individual i, we assume that conditional on C, the actual treatment level T is independent of each of the potential outcomes: $Y(t) \perp T \mid \mathbf{C}, \forall t$

C.2 Evaluation of the tree consistency

We evaluated the consistency of our clustering across subsamples using an Adjusted Rand Index [141]. Given a set of n elements $S = \{o_1, \ldots, o_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and $Y = \{Y_1, \ldots, Y_s\}$, a partition of S into s subsets, define the following:

- a, the number of pairs of elements in S that are in the same subset in X and in the same subset in Y.
- b, the number of pairs of elements in S that are in different subsets in X and in different subsets in Y.
- c, the number of pairs of elements in S that are in the same subset in X and in different subsets in Y.
- d, the number of pairs of elements in S that are in different subsets in X and in the same subset in Y.

The Rand index, R, is:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Intuitively, $a+b$ can be considered as the number of agreements between X and Y and $c+d$ as the number of disagreements between X and Y. The adjusted Rand index is the corrected-for-chance version of the Rand index [238].

C.3 Proof of convergence

C.3.1 Proof of ASMD Decrease with Multivariate Gaussian X

Lemma 4 (Multivariate Normal Case). Let $\mathbf{X} = (X_1, \ldots, X_p)$ be multivariate normal random variables and T a binary treatment indicator. If the following conditions hold:

1. $|\rho_{ij}| \leq \delta \leq \frac{1-K}{CM^2}$ for all $i \neq j$, where C is a constant dependent on the split point
2. $\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}$ for all $i \neq j$, with $K < 1$

3. $\frac{\sigma_i}{\sigma_j} \leq M$ for all $i \neq j$, with $M \geq 1$

4. $\text{Var}(X_i|T=1) = \text{Var}(X_i|T=0) = \sigma_i^2$ for all i

Then, after splitting on the variable X_j with the highest ASMD, the maximum ASMD in the resulting child nodes will not exceed the original maximum ASMD in the parent node. Put differently, the ASMD imbalance in children nodes either decreases or remains the same.
Note: While the proof assumes equal variances for simplicity, similar reasoning can be extended to cases where variances differ between treatment groups.

Problem Statement

Given:

- A set of continuous random variables $X = (X_1, X_2, \ldots, X_p)$.
- A binary variable T.
- X follows a multivariate Gaussian (normal) distribution.
- The Absolute Standardized Mean Difference (ASMD) for variable X_i is defined as:

$$\text{ASMD}_i = \frac{|\mathbb{E}[X_i \mid T=1] - \mathbb{E}[X_i \mid T=0]|}{\sqrt{\text{Var}(X_i \mid T=1) + \text{Var}(X_i \mid T=0)}}$$

Procedure:

- Identify X_j with the maximum ASMD in the parent node.
- Split the data on X_j at the value x_j that yields the lowest p -value from Fisher's exact test.

Goal: Prove that under certain assumptions, the maximum ASMD over all variables X_i in the child nodes does not exceed the original maximum ASMD in the parent node.

Assumptions

1. **Bounded Correlation Between Variables:** The correlations between X_j and other variables X_i (for $i \neq j$) are bounded:

$$|\rho_{ij}| \leq \delta, \quad \text{for all } i \neq j$$

where δ is a small positive constant (e.g., $\delta \leq 0.1$).

2. **Initial ASMDs of other variables:** The initial ASMDs of other variables X_i are lower than that of X_j :

$$\mathrm{ASMD}_i^{\text{parent}} \leq K \cdot \mathrm{ASMD}_j^{\text{parent}}, \text{ for all } i \neq j$$

where K is a constant less than 1 (e.g., $K \leq 0.5$).

3. **Equal Variances Across Groups:** The variances of X_i conditional on T are equal:

$$\operatorname{Var}(X_i \mid T=1) = \operatorname{Var}(X_i \mid T=0) = \sigma_i^2$$

4. **Bounded Ratio of Standard Deviations:** There exists a constant $M \geq 1$ such that:

$$\frac{\sigma_i}{\sigma_j} \leq M, \quad \text{for all } i \neq j$$

5. **Linearity and Normality:** Utilizes properties of the multivariate normal distribution, implying linear relationships between variables.

Proof Outline

1. Express ASMDs in terms of parameters.
2. Derive the ASMD of X_i in the child nodes.
3. Bound the change in ASMD of X_i Using δ.
4. Demonstrate that under the given assumptions, the ASMD of X_i in the child nodes does not exceed the original maximum ASMD.

Detailed Proof 1. Express ASMDs in Terms of Parameters
For Variable X_i : In the parent node:

$$\mathrm{ASMD}_i^{\text{parent}} = \frac{|\mu_{i1} - \mu_{i0}|}{\sqrt{2}\sigma_i} = \frac{|\Delta\mu_i|}{\sqrt{2}\sigma_i}$$

where: $\mu_i = \mathbb{E}[X_i \mid T=t]$ and $\Delta\mu_i = \mu_{i1} - \mu_{i0}$
Similarly for variable X_j :

$$\mathrm{ASMD}_j^{\text{parent}} = \frac{|\Delta\mu_j|}{\sqrt{2}\sigma_j}$$

2. Derive the ASMD of X_i in the child nodes

After splitting on X_j at x_j, consider one of the child nodes (e.g., $X_j \leq x_j$). Conditional Mean of X_i for $T=t$:

$$\mathbb{E}\left[X_i \mid X_j \leq x_j, T=t\right]=\mu_{it}+\rho_{ij} \frac{\sigma_i}{\sigma_j}\left(\mathbb{E}\left[X_j \mid X_j \leq x_j, T=t\right]-\mu_{jt}\right)$$

Difference in Conditional Means between $T=1$ and $T=0$:

$$\Delta \mu_i^{\text{child}}=\Delta \mu_i+\rho_{ij} \frac{\sigma_i}{\sigma_j}\left(\Delta \mu_j^{\text{trunc },1}-\Delta \mu_j^{\text{trunc },0}\right)$$

where: $\Delta \mu_j^{\text{trunc },t}=\mathbb{E}\left[X_j \mid X_j \leq x_j, T=t\right]-\mu_{jt}$

3. Bound the change in ASMD of X_i using δ

Absolute difference in adjustments:

$$\left|\Delta \mu_i^{\text{child}}-\Delta \mu_i\right|=\left|\rho_{ij} \frac{\sigma_i}{\sigma_j}\left(\Delta \mu_j^{\text{trunc, } 1}-\Delta \mu_j^{\text{trunc },0}\right)\right|$$

Since $\left|\rho_{ij}\right| \leq \delta$ and $\frac{\sigma_i}{\sigma_j} \leq M$,

$$\left|\Delta \mu_i^{\text{child}}-\Delta \mu_i\right| \leq \delta \cdot \frac{\sigma_i}{\sigma_j}\left|\Delta \mu_j^{\text{trunc, },1}-\Delta \mu_j^{\text{trunc },0}\right| \leq \delta \cdot M\left|\Delta \mu_j^{\text{trunc },1}-\Delta \mu_j^{\text{trunc },0}\right|$$

Bounding $\left|\Delta \mu_j^{\text{trunc, } 1}-\Delta \mu_j^{\text{trunc, },0}\right|$, we assume a constant C such that:

$$\left|\Delta \mu_j^{\text{trunc, } 1}-\Delta \mu_j^{\text{trunc, } 0}\right| \leq C\left|\Delta \mu_j\right|$$

Note that this term is proportional to $\left|\Delta \mu_j\right|$ **because X is multivariate Gaussian.** (proof in the following subsection)

Therefore:

$$\left|\Delta \mu_i^{\text{child}}-\Delta \mu_i\right| \leq C M \delta\left|\Delta \mu_j\right|$$

ASMD of X_i in the child node:

$$\mathrm{ASMD}_i^{\text{child}}=\frac{\left|\Delta \mu_i^{\text{child}}\right|}{\sqrt{2} \sigma_i} \leq \frac{\left|\Delta \mu_i\right|+C M \delta\left|\Delta \mu_j\right|}{\sqrt{2} \sigma_i}$$

or

$$\mathrm{ASMD}_i^{\text{child}} \leq \mathrm{ASMD}_i^{\text{parent}}+\frac{C M \delta\left|\Delta \mu_j\right|}{\sqrt{2} \sigma_i}$$

4. Demonstrate that $\mathrm{ASMD}_i^{\textbf{child}} \leq \mathrm{ASMD}_j^{\textbf{parent}}$

Express $\left|\Delta \mu_j\right|$ in Terms of $\mathrm{ASMD}_j^{\text{parent}}$

$$\left|\Delta\mu_j\right| = \sqrt{2}\sigma_j \cdot \mathrm{ASMD}_j^{\mathrm{parent}}$$

Substitute back into the ASMD of X_i :

$$\mathrm{ASMD}_i^{\mathrm{child}} \leq \mathrm{ASMD}_i^{\mathrm{parent}} + CM\delta \cdot \frac{\sqrt{2}\sigma_j \cdot \mathrm{ASMD}_j^{\mathrm{parent}}}{\sqrt{2}\sigma_i} = \mathrm{ASMD}_i^{\mathrm{parent}} + \frac{CM\delta\sigma_j}{\sigma_i} \cdot \mathrm{ASMD}_j^{\mathrm{parent}}$$

Using $\frac{\sigma_i}{\sigma_j} \leq M$:

$$\mathrm{ASMD}_i^{\mathrm{child}} \leq \mathrm{ASMD}_i^{\mathrm{parent}} + CM^2\delta \cdot \mathrm{ASMD}_j^{\mathrm{parent}}$$

To ensure $\mathrm{ASMD}_i^{\mathrm{child}} \leq \mathrm{ASMD}_j^{\mathrm{parent}}$ we require:

$$\mathrm{ASMD}_i^{\mathrm{parent}} + CM^2\delta \cdot \mathrm{ASMD}_j^{\mathrm{parent}} \leq \mathrm{ASMD}_j^{\mathrm{parent}}$$

Subtract $\mathrm{ASMD}_i^{\mathrm{parent}}$ from both sides:

$$CM^2\delta \cdot \mathrm{ASMD}_j^{\mathrm{parent}} \leq \mathrm{ASMD}_j^{\mathrm{parent}} - \mathrm{ASMD}_i^{\mathrm{parent}}$$

Divide both sides by $\mathrm{ASMD}_j^{\mathrm{parent}}$:

$$CM^2\delta \leq 1 - \frac{\mathrm{ASMD}_i^{\mathrm{parent}}}{\mathrm{ASMD}_j^{\mathrm{parent}}}$$

Using the assumption $\mathrm{ASMD}_i^{\mathrm{parent}} \leq K \cdot \mathrm{ASMD}_j^{\mathrm{parent}}$:

$$CM^2\delta \leq 1 - K$$

Therefore, to satisfy the inequality:

$$\delta \leq \frac{1-K}{CM^2}$$

Since C is a constant dependent on the truncation and properties of the normal distribution, and $K < 1$, this inequality provides an upper bound on δ.

Under the given assumptions, this recursive process will lead to a decrease in ASMD imbalance as we continue to split.

Note: The proof assumes equal variances for simplicity, but similar reasoning can be extended when variances differ.

C.3.2 Proof of ASMD Decrease with Mixed Data Types X Using Gaussian Copula

Lemma 5 (Mixed Data Types Case using Gaussian Copula). Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector of mixed continuous and categorical variables, T a binary treatment indicator, and assume a Gaussian copula model with latent variables $\mathbf{Z}$. If the following conditions hold:

1. $|\rho_{ij}| \leq \delta \leq \frac{1-K}{M^2 C}$ for all $i \neq j$, where C is a constant dependent on the split point
2. $\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}$ for all $i \neq j$, with $K < 1$
3. For continuous variables, $\frac{\sigma_i}{\sigma_j} \leq M$ for all $i \neq j$, with $M \geq 1$
4. For continuous variables, $\text{Var}(X_i | T = 1) = \text{Var}(X_i | T = 0) = \sigma_i^2$
5. For categorical variables, the thresholds determining categories are the same across treatment groups

Then, after splitting on the variable X_j with the highest ASMD, the maximum ASMD in the resulting child nodes will not exceed the original maximum ASMD in the parent node. Put differently, the ASMD imbalance in children nodes either decreases or remains the same.
Note: While the proof assumes equal variances for simplicity, similar reasoning can be extended to cases where variances differ between treatment groups.

Assumptions

1. **Variables and Treatment:**
   - Let $X = (X_1, X_2, \ldots, X_p)$ be a vector of variables consisting of both continuous and categorical variables.
   - Let T be a binary treatment indicator ($T = 0$ or 1).
2. **Gaussian Copula Model:**
   - There exists a latent variable vector $Z = (Z_1, Z_2, \ldots, Z_p)$ such that:

$$(Z, T) \sim \mathcal{N}_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

   where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix capturing the dependence structure between variables and the treatment.

- The observed variables X_i are obtained by applying the inverse of their marginal cumulative distribution functions to the standard normal cumulative distribution of the latent variables:
$$X_i = F_i^{-1}(\Phi(Z_i)),$$
where F_i^{-1} is the inverse CDF (quantile function) of X_i, and Φ is the standard normal CDF.

3. **Latent Variable Representation for Categorical Variables:**
    - For each categorical variable X_i, there exists a latent continuous variable X_i^*.
    - The observed categorical variable X_i is determined by thresholding X_i^*:
    $$X_i = k \quad \text{if} \quad X_i^* \in (\tau_{i,k-1}, \tau_{i,k}],$$
    where $\{\tau_{i,k}\}$ are threshold parameters with $-\infty = \tau_{i,0} < \tau_{i,1} < \ldots < \tau_{i,K_i} = \infty$.

4. **Bounded Correlations:**
    - The correlations between the splitting variable X_j and other variables X_i are bounded:
    $$|\rho_{ij}| \leq \delta, \quad \forall i \neq j,$$
    where δ is a small positive constant.

5. **Initial ASMDs:**
    - The initial Absolute Standardized Mean Differences (ASMDs) of other variables X_i are significantly lower than that of X_j:
    $$\text{ASMD}_i^{\text{parent}} \leq K \cdot \text{ASMD}_j^{\text{parent}}, \quad \text{with} \quad K < 1.$$

6. **Standard Deviations:**
    - For continuous variables, the standard deviations satisfy:
    $$\frac{\sigma_i}{\sigma_j} \leq M, \quad \forall i \neq j,$$
    where $M \geq 1$ is a constant.

7. **Equality of Variances Across Treatment Groups:**
    - For continuous variables:
    $$\text{Var}(X_i \mid T = 1) = \text{Var}(X_i \mid T = 0) = \sigma_i^2.$$

8. **Thresholds for Categorical Variables:**

    - The thresholds $\{\tau_{i,k}\}$ used to determine the observed categories from latent variables are the same across treatment groups.

—

Proof **Step 1: Express ASMDs in Terms of Latent Variables**
Continuous Variables:
For a continuous variable X_i, the ASMD in the parent node is:

$$\text{ASMD}_i^{\text{parent}} = \frac{|\mu_{i1} - \mu_{i0}|}{\sqrt{2}\sigma_i},$$

where $\mu_{it} = \mathbb{E}[X_i \mid T = t]$.
Categorical Variables:
For a categorical variable X_i, modeled via a latent variable X_i^*:

- The probability that X_i takes category k in group $T = t$ is:

$$p_{it,k} = \mathbb{P}(X_i = k \mid T = t) = \Phi\left(\frac{\tau_{i,k} - \mu_{it}^*}{\sigma_i^*}\right) - \Phi\left(\frac{\tau_{i,k-1} - \mu_{it}^*}{\sigma_i^*}\right).$$

- The ASMD for category k can be defined as:

$$\text{ASMD}_{i,k}^{\text{parent}} = \frac{2(p_{i1,k} - p_{i0,k})}{\sqrt{p_{i1,k}(1 - p_{i1,k}) + p_{i0,k}(1 - p_{i0,k})}}.$$

Step 2: Introduction of Constant C in the Gaussian Copula
In the context of the Gaussian copula, the dependence between variables is captured by the covariance matrix Σ. When we split on variable X_j, we consider the conditional distribution of other variables given X_j.
For Continuous Variables:
The conditional expectation of X_i given $X_j \leq x_j$ and $T = t$ is:

$$\mathbb{E}[X_i \mid X_j \leq x_j, T = t] = \mu_{it} + \rho_{ij}\frac{\sigma_i}{\sigma_j}\left(\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}\right).$$

The term $\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt}$ can be expressed using properties of the truncated normal distribution:

$$\mathbb{E}[X_j \mid X_j \leq x_j, T = t] - \mu_{jt} = -\sigma_j\lambda_t,$$

where λ_t is defined as:

$$\lambda_t = \frac{\phi(z_t)}{\Phi(z_t)},$$

with $z_t = \frac{x_j - \mu_{jt}}{\sigma_j}$, and ϕ and Φ are the standard normal PDF and CDF, respectively.
Difference Between Treatment Groups:
The difference in these terms between the two treatment groups is:

$$\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0} = -\sigma_j(\lambda_1 - \lambda_0).$$

Using a Taylor expansion around $z = \frac{x_j - \mu_j}{\sigma_j}$, where $\mu_j = \frac{\mu_{j1}+\mu_{j0}}{2}$, we approximate λ_t as:

$$\lambda_t \approx \lambda - \lambda' \left((-1)^{t+1} \frac{\Delta\mu_j}{2\sigma_j} \right),$$

where:
$\lambda = \frac{\phi(z)}{\Phi(z)}$, and $\lambda' = \frac{d}{dz}\left(\frac{\phi(z)}{\Phi(z)}\right) = -\lambda(z+\lambda)$.
Therefore, the difference $\lambda_1 - \lambda_0$ is:

$$\lambda_1 - \lambda_0 \approx \lambda' \left(\frac{\Delta\mu_j}{\sigma_j} \right).$$

Substituting back, we have:

$$|\Delta\mu_j^{\text{trunc},1} - \Delta\mu_j^{\text{trunc},0}| = \left|-\sigma_j(\lambda_1 - \lambda_0)\right| = |\lambda'||\Delta\mu_j|.$$

Define the constant C as:

$$C = |\lambda'| = \lambda|z+\lambda|.$$

This constant C depends on the truncation point x_j and captures the effect of truncation on the mean differences.
Result:
Under the given assumptions, and per the previous result in the continuous (multivariate Gaussian) case we have:

$$\delta \leq \frac{1-K}{M^2C},$$

we have that the ASMDs of both continuous and categorical variables in the child nodes after splitting do not exceed $\text{ASMD}_j^{\text{parent}}$:

$$\text{ASMD}_i^{\text{child}} \leq \text{ASMD}_j^{\text{parent}}, \quad \forall i \neq j.$$

Calculation of C in the Multivariate Gaussian and Gaussian Copula Cases

Multivariate Gaussian Case Derivation of C:
Consider a continuous variable X_j that follows a normal distribution within each treatment group $T = t$:

$$X_j \mid T = t \sim \mathcal{N}(\mu_{jt}, \sigma_j^2).$$

When we split on X_j at a threshold x_j, the truncated means for each treatment group are:

$$\mu_{jt}^{\text{trunc}} = \mathbb{E}[X_j \mid X_j \leq x_j, T = t] = \mu_{jt} - \sigma_j \lambda_t,$$

where λ_t is the inverse Mills ratio:

$$\lambda_t = \frac{\phi(z_t)}{\Phi(z_t)},$$

with $z_t = \dfrac{x_j - \mu_{jt}}{\sigma_j}$, $\phi(\cdot)$ is the standard normal PDF, and $\Phi(\cdot)$ is the standard normal CDF. The difference in truncated means between treatment groups is:

$$\Delta\mu_j^{\text{trunc}} = \mu_{j1}^{\text{trunc}} - \mu_{j0}^{\text{trunc}} = (\mu_{j1} - \sigma_j \lambda_1) - (\mu_{j0} - \sigma_j \lambda_0) = \Delta\mu_j - \sigma_j(\lambda_1 - \lambda_0),$$

where $\Delta\mu_j = \mu_{j1} - \mu_{j0}$.
To approximate λ_t, we perform a first-order Taylor expansion around $\bar{\mu}_j = \dfrac{\mu_{j1} + \mu_{j0}}{2}$:

$$\lambda_t \approx \lambda - \lambda' \left((-1)^{t+1} \frac{\Delta\mu_j}{2\sigma_j} \right),$$

where:

- $z = \dfrac{x_j - \bar{\mu}_j}{\sigma_j}$,
- $\lambda = \dfrac{\phi(z)}{\Phi(z)}$,
- $\lambda' = \dfrac{d}{dz} \left(\dfrac{\phi(z)}{\Phi(z)} \right) = -\lambda(z + \lambda$

Then, the difference $\lambda_1 - \lambda_0$ is:

$$\lambda_1 - \lambda_0 \approx \lambda' \left(\frac{\Delta\mu_j}{\sigma_j} \right).$$

Substituting back:

$$\Delta\mu_j^{\text{trunc}} = \Delta\mu_j - \sigma_j \left(\lambda' \frac{\Delta\mu_j}{\sigma_j} \right) = \Delta\mu_j - \lambda' \Delta\mu_j = (1 - \lambda')\Delta\mu_j.$$

The magnitude of the change in the truncated mean difference is:

$$\left|\Delta\mu_j^{\text{trunc}} - \Delta\mu_j\right| = \left|-\lambda' \Delta\mu_j\right| = |\lambda'| \left|\Delta\mu_j\right|.$$

Define the Constant C:

$$C = |\lambda'| = \lambda |z + \lambda|.$$

Gaussian Copula Case Derivation of C:

In the Gaussian copula framework, the dependence structure between variables is modeled using a multivariate normal distribution of latent variables Z:

$$Z = (Z_1, Z_2, \ldots, Z_p) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ is the correlation matrix. The observed variables X_j are obtained by transforming the latent variables using their marginal distributions:

$$X_j = F_j^{-1}\left(\Phi(Z_j)\right),$$

where F_j^{-1} is the inverse CDF of X_j.

When splitting on X_j, we consider the conditional distribution of Z_j given $Z_j \leq z_j$. The truncated mean of Z_j is:

$$\mu_{Z_j}^{\text{trunc}} = \mathbb{E}[Z_j \mid Z_j \leq z_j] = -\lambda,$$

where $\lambda = \dfrac{\phi(z_j)}{\Phi(z_j)}$

The difference in truncated means between treatment groups (assuming different thresholds for Z_j) is similarly approximated. Using Taylor expansion as in the multivariate Gaussian case, we arrive at:

$$\left|\Delta\mu_{Z_j}^{\text{trunc}} - \Delta\mu_{Z_j}\right| = |\lambda'| \left|\Delta\mu_{Z_j}\right|,$$

where $\Delta\mu_{Z_j}$ is the difference in means of Z_j between treatment groups, and:

$$C = |\lambda'| = \lambda |z_j + \lambda|.$$

Since the observed X_j is a function of Z_j through F_j^{-1}, the change in $\Delta\mu_j^{\text{trunc}}$ can be related back to $\Delta\mu_{Z_j}^{\text{trunc}}$, and C serves the same role in bounding the change.

C.4 Further related work

Causal inference provides a wide range of methods for estimating causal effect from data with unbalanced treatment allocation. In balancing methods such as matching or weighting methods, the data is pre-processed to create subgroups with lower treatment imbalance or "natural experiments".

C.4.1 Effect estimation

Matching *Matching* methods consist of clustering similar units from the treatment and control groups to reduce imbalance. In general, a matching procedure generates weights w_{ij} denoting the assignment of one or many control units j to a treated unit i ([239], Chapter 5). Exact matching only assigns control units to treatment units with the exact same set of covariate values. But typically, matched control units j are chosen based on a nearest-neighbors search according to some distance metric. However, matching procedures induce a bias-variance trade-off as discarding unmatched samples reduces estimation error at the cost of increased variance. However, all matching methods suffer from the curse of dimensionality, making them impractical in high-dimensional datasets. Exact matching and coarsened exact matching [137] find exponentially fewer matches as the input dimension grows [240]. Alternative methods include Propensity score matching [241], where distance is computed from an estimate of the propensity score $P(T = 1 \mid X = \mathbf{x})$, and Mahalanobis distance matching [135] (see more details below in 5.4.1). However, compression into a single dimension can lead to highly unrelated matches with very different characteristics in the original covariate space, and can ultimately increase estimation bias [242]. *Clivio et. al* overcome this issue by developing a multivariate balancing score to perform matching for high-dimensional causal inference. Nevertheless, this approach is not interpretable.

Weighting methods An alternative to matching are *weighting* methods, where sample weights are estimated, generalizing the problem formulation of matching. In Inverse Probability Weighting (IPW) [46] which is the most popular alternative in that category, the samples are weighted according to their *propensity* score, i.e. the estimated probability of treatment conditional on their covariates.

Outcome models *Outcome* models estimate the causal effect from regression models where both treatment and covariates act as predictors of the outcome. These regressions can be fitted through various methods like linear regression [7], neural networks [121, 122], or tree-based models [40]. Common alternatives include Doubly Robust estimators [243], Double Debiased Machine Learning [244] and metalearners such as the T-learner and

X-learner [42]. These methods have the advantage of being very data-adaptive. More particularly, the Causal Tree approach [40] builds on regression tree methods, and splits the data to optimize for goodness of fit in treatment effects. Causal Tree separates the training dataset into two subsamples: a splitting subsample and an estimating subsample. The splitting subsample is used to build a causal tree while the estimating subsample is used to generate unbiased conditional treatment effect estimates. This procedure is called "honest estimation" and is anticipated to avoid overfitting. However, most outcome models are not interpretable.

Representation learning Representation learning has emerged as a key approach for causal effect estimation, especially in observational settings where confounding is a significant challenge. By learning low-dimensional representations of covariates, these methods aim to reduce bias and improve the estimation of treatment effects. For instance, neural networks have been used to capture complex relationships between variables, as shown in [**?**]. Optimal transport methods have further enhanced these estimates by ensuring balance between treatment groups [**?**]. Additionally, decomposed representations, which separate confounding information from treatment-specific effects, allow for more accurate estimates [169], while disentangled representations help isolate factors affecting potential outcomes [**?**]. However, by definition using representation learning prevents any interpretability of the effect estimation procedure.

C.4.2 Positivity violations

Causal inference is only possible under the *positivity* assumption, which requires covariate distributions to overlap between treatment arms. Thus, positivity violations (also referred to as no overlap) occur when certain subgroups in a sample do not receive one of the treatments of interest or receive it too rarely [123]. Overlap is essential as it guarantees data-driven outcome extrapolation across treatment groups. Having no common support means there are subjects in one group with no counterparts from the other group, and, therefore, no reliable way to pool information on their outcome had they been in the other group. Non-violating samples are thus the only ones for which we can guarantee some validity of the inferred causal effect.

There are three common ways to characterize positivity. The most common one consists in estimating propensity scores and excluding the samples associated with extreme values (known as "trimming") [124]. The threshold for propensity scores can be set arbitrarily or dynamically [37]. However, since samples are excluded on the basis of their propensity scores and not their covariate values, these methods lack interpretability about the excluded subjects and how it may affect the target population on which we can generalize the inference. Thus, other methods have been developed to overcome this challenge by characterizing the propensity-based exclusion [125, 126, 127]. Lastly, the third way tries to characterize the overlap from covariates and treatment assignment directly, without going through the intermediate propensity score e.g. PositiviTree [123]. In PositiviTree, a decision tree classifier is fitted to predict treatment allocation. In contrast to their approach, BICause Tree implements a tailor-made optimization function where splits are chosen to maximize balancing in the resulting sub-population, whereas PositiviTree uses off-the-shelf decision trees maximizing separation. Ultimately, the above-mentioned methods for positivity identification and characterization are model agnostic. In our model, BICauseTree, positivity identification *and characterization* are inherently integrated into the model, and effect estimation comes with a built-in interpretable abstention prediction mechanism.

C.4.3 Further comparison with Causal Tree

Causal Trees (CT) [40] are another tree-based model for causal inference that (i) leverages the inherent interpretability of decision trees, and (ii) has a custom objective function for recursively splitting the data. Although both utilize decision trees, BICauseTree and CT serve distinct purposes. BICauseTree splits are optimized for balancing treatment *allocation* while CT splits are optimized for balancing treatment *effect*, under assumed exchangeability. In other words, CT *assumes* exchangeability while BICauseTree *finds* exchangeability. As such, our approach is only suited for ATE estimation while CT is better suited for Conditional Average Treatment Effect estimation [40]. Furthermore, in practice, causal effects are often averaged over multiple trees into a Causal *Forest* [41, 245] that is no longer interpretable, and users are encouraged to use post-hoc explanation methods [246].

C.5 Experimental details: synthetic datasets

C.5.1 Natural experiment dataset

For the natural experiment dataset, we consider a Death outcome D, a binary treatment of interest T and two covariates such that $X = (S, A)$ with S the sex and A the continuous age such that:

$$S \sim \text{Bernoulli}(0.5)$$

$$A \sim \text{Normal}\left(50, 20^2\right)$$

The sample size was chosen to be $N = 20,000$.

We defined four sub-populations, each constituting a natural experiment, with a different propensity distribution $P(T=1|X=x)$:

$$\Pr(T=1|S=1,A\geq 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.5,0.1^2\right)$$
$$\Pr(T=1|S=1,A<50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.3,0.1^2\right)$$
$$\Pr(T=1|S=0,A\geq 50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.1,0.1^2\right)$$
$$\Pr(T=1|S=0,A<50) \sim \text{TruncatedNormal}_{[0,1]}\left(0.4,0.1^2\right)$$

Individual treatment propensities were sampled from the corresponding distributions above and observed treatment values were sampled from a Bernoulli distribution parameterised with the individual propensities. The outcome probabilities were not modelled as a distribution. Instead, observed outcome values were sampled directly from a Bernoulli distribution parameterised with a constant value that depended on both X and T.

- $\Pr(Y|T=1,S=1,A\geq 50) \sim \mathscr{B}(0.1)$
- $\Pr(Y|T=1,S=1,A<50) \sim \mathscr{B}(0.2)$
- $\Pr(Y|T=1,S=0,A\geq 50) \sim \mathscr{B}(0.4)$
- $\Pr(Y|T=1,S=0,A<50) \sim \mathscr{B}(0.15)$
- $\Pr(Y|T=0,S=1,A\geq 50) \sim \mathscr{B}(0.2)$
- $\Pr(Y|T=0,S=1,A<50) \sim \mathscr{B}(0.4)$
- $\Pr(Y|T=0,S=0,A\geq 50) \sim \mathscr{B}(0.8)$
- $\Pr(Y|T=0,S=0,A<50) \sim \mathscr{B}(0.3)$

No positivity violation was modelled in this experiment.

C.5.1.1 Positivity violations dataset

For this second synthetic experiment, we build a dataset with two positivity violating subgroups. Let us consider a synthetic example of a dataset with a Death outcome D, a binary treatment of interest T and three binary covariates –sex, cancer and arrhythmia– such that $X=(S,C,A)$. For the marginal distributions, we set:

$$S \sim \text{Ber}(0.5)$$
$$C \sim \text{Ber}(0.3)$$
$$A \sim \text{Ber}(0.1)$$

The sample size was chosen to be $N=20,000$.

- $\Pr(T=1|S=1,C=1,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(1.00,0.02^2\right)$

- $\Pr(T=1|S=1,C=0,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.32, 0.10^2\right)$
- $\Pr(T=1|S=1,C=1,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.12, 0.10^2\right)$
- $\Pr(T=1|S=1,C=0,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.42, 0.10^2\right)$
- $\Pr(T=1|S=0,C=1,A=0) \sim \text{TruncatedNormal}_{[0,1]}\left(0.17, 0.10^2\right)$
- $\Pr(T=1|S=0,C=1,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.30, 0.10^2\right)$
- $\Pr(T=1|S=0,C=0,A=1) \sim \text{TruncatedNormal}_{[0,1]}\left(0.24, 0.10^2\right)$
- $\Pr(T=1|S=0,C=0,A=0) \sim \text{TruncatedNormal}_{[0,1]}(0.00, 0.02)$

The observed outcome values were sampled from a Bernoulli distribution parameterised with a constant value that depended on both X and T. For the treated:

- $\Pr(Y|T=1,S=1,C=1,A=1) \sim \mathscr{B}(0.13)$
- $\Pr(Y|T=1,S=1,C=1,A=0) \sim \mathscr{B}(0.08)$
- $\Pr(Y|T=1,S=1,C=0,A=1) \sim \mathscr{B}(0.21)$
- $\Pr(Y|T=1,S=1,C=0,A=0) \sim \mathscr{B}(0.1)$
- $\Pr(Y|T=1,S=0,C=1,A=1) \sim \mathscr{B}(0.36)$
- $\Pr(Y|T=1,S=0,C=1,A=0) \sim \mathscr{B}(0.29)$
- $\Pr(Y|T=1,S=0,C=0,A=1) \sim \mathscr{B}(0.24)$
- $\Pr(Y|T=1,S=0,C=0,A=0) \sim \mathscr{B}(0.09)$

For the untreated:

- $\Pr(Y|T=0,S=1,C=1,A=1) \sim \mathscr{B}(0.31)$
- $\Pr(Y|T=0,S=1,C=1,A=0) \sim \mathscr{B}(0.4)$
- $\Pr(Y|T=0,S=1,C=0,A=1) \sim \mathscr{B}(0.29)$
- $\Pr(Y|T=0,S=1,C=0,A=0) \sim \mathscr{B}(0.45)$
- $\Pr(Y|T=0,S=0,C=1,A=1) \sim \mathscr{B}(0.4)$
- $\Pr(Y|T=0,S=0,C=1,A=0) \sim \mathscr{B}(0.51)$
- $\Pr(Y|T=0,S=0,C=0,A=1) \sim \mathscr{B}(0.43)$
- $\Pr(Y|T=0,S=0,C=0,A=0) \sim \mathscr{B}(0.73)$

C.5.2 Further experiment results: synthetic experiments

C.5.2.1 Natural experiment dataset

Adding Coarsened Exact Matching Coarsened Exact Matching (CEM) [**?**] is a matching method for causal inference that temporarily coarsens covariates, matches treated and control units exactly on these coarsened values, and then retains the original (uncoarsened) values for analysis, reducing model dependence and improving balance. On the natural experiment, CEM exhibits performance comparable to Mahalanobis Matching (MM).

Method	Sample	Mean Bias	Variance
BICause Tree	Test	0.012	0.004
BICause Tree	Train	0.009	0.004
IPW (LR)	Test	0.060	0.012
IPW (LR)	Train	0.058	0.010
IPW (GBT)	Test	0.008	0.006
IPW (GBT)	Train	0.007	0.005
Marginal	Test	0.072	0.011
Marginal	Train	0.067	0.010
MM	Test	0.012	0.018
MM	Train	0.007	0.012
CEM	Test	0.013	0.018
CEM	Train	0.011	0.012

Table C1: Mean estimation bias and estimation variance for the natural experiment dataset. MM: Mahalanobis Matching, IPW: Inverse Propensity Score Weighting, CEM: Coarsened Exact Matching

The importance of selecting candidate features by max ASMD The heuristic for choosing the feature with the maximum absolute standardized mean difference (ASMD) among all features to split on can be empirically justified. We can do that by comparing two BICauseTrees: one that selects the feature with maximal ASMD at each recursive iteration (the model used throughout this work), to a second tree whose only difference is choosing candidate features randomly. Figure C1 demonstrates how this design decision contributes to a reduction in estimation bias.

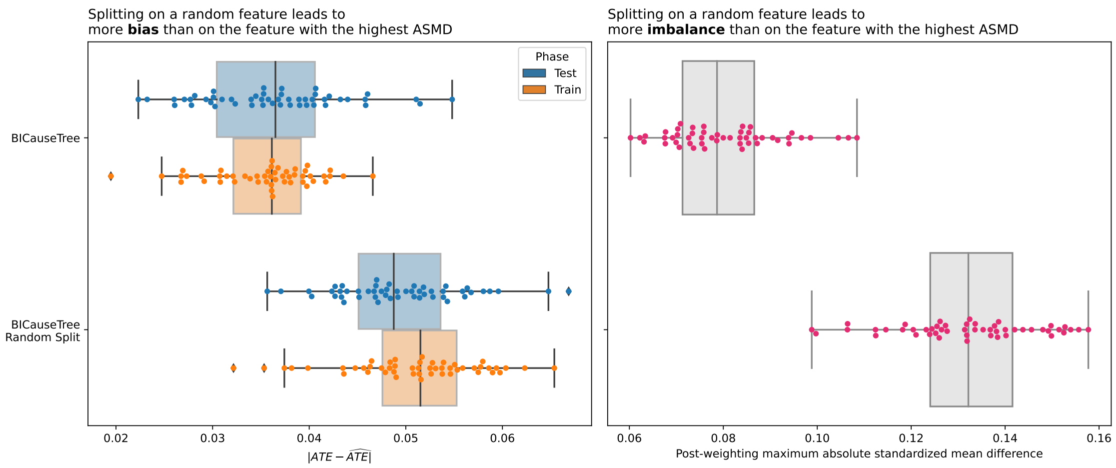


Figure C1: Comparison of estimation bias (left, absolute difference) and imbalance (right, maximum absolute standardized mean difference [ASMD]) using two flavors of BICauseTree: First, the original version from the manuscript, wherein each recursion the feature to split on is chosen by the one with maximal ASMD. And second, a variant where features are selected at random, with all other hyperparameters held constant. We see that selecting the most imbalanced feature for stratification leads to better balancing and, more importantly, better estimation - justifying the rationale for selecting features based on the highest ASMD.

Additionally, adjusting for the maximal ASMD can be justified in simplistic scenarios. Consider the following data generating process:

$$Y = A + X_1 + X_2$$

$$A \sim Bernoulli(0.5)$$

$$X_1, X_2 \sim Normal(0,1)$$

$$X_2|_{A=1} \sim Normal(1,1)$$

Where X_1 is balanced (independent of A), but X_2 is not since the mean of the treated samples is shifted one unit upward.

We can then consider four simple linear models:

1) Null-model: $Y \sim 1 + A$.

2) Full-model: $Y \sim 1 + A + X_1 + X_2$.

3) X_1-model: $Y \sim 1 + A + X_1$.

4) X_2-model: $Y \sim 1 + A + X_2$.

And observe that the X_2-model recovers the true effect much better than the X_1-model (see Figure C2).

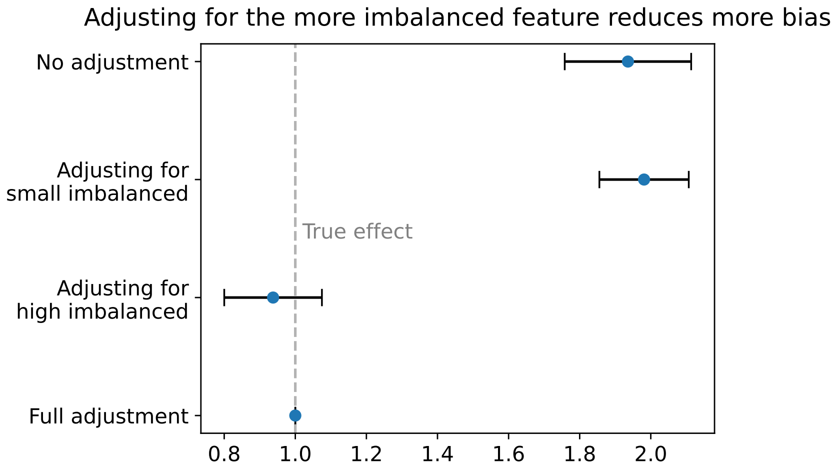


Figure C2: Adjusting for the more imbalanced predictor reduces more estimation bias of the treatment effect. The data contains a treatment assignment and two predictor variables – one balanced across treatment groups and one imbalanced. We examine four models, all adjusting for the treatment assignment but differing in their adjustment of the predictors. Adjusting for the more imbalanced predictor almost retrieves the treatment effect perfectly (dashed grey line), while adjusting the less imbalanced predictor still suffers from large residual confounding.

Note that this implicitly assumes all predictors have an equal association with the outcome. Therefore, as suggested in the main text, future work can create a combine score of ASMD and predictor-outcome association. One such approach can be as follows.

1. We calculate the ASMDs, denoted s.

2. We then calculate any variable importance measure, preferably fitted jointly using all predictors and treatment assignment, as long as its values ranges in the positives. Denote it as q.
   Examples can be the absolute coefficients for regression models or permutation importance for tree-based models.

3. We then normalize both s and q, separately so each sum to 1, depicting relative imbalance and relative predictor-outcome importance among predictors.

4. Finally, we use the combined measurement $s \cdot q$ to select the predictor to split on in each iteration.

Examining $s \cdot q$ it is easy to see that if a variable j is balanced (near-zero s_j) or unrelated to the outcome (near-zero q_j), the combined score will be small and less likely to be chosen for splitting. The predictor j itself is also less likely to cause bias because it is either balanced, or, if imbalanced, affects the outcome very little.

Applicability in high-dimensional data Being based on decision tree logic allows BICauseTree to scale properly for high-dimensional data. BICauseTree, like decision trees, scans all the features before deciding which to split on. This allows it to focus on the most biasing feature at each partition, regardless of their number, enabling them to find natural experiments in noisy high-dimensional data. To demonstrate that, we augmented our natural experiment dataset with an increasing number of noisy features. Figure C3 shows that the error of BICauseTree's ATE estimation stays consistent regardless of the number of added features.

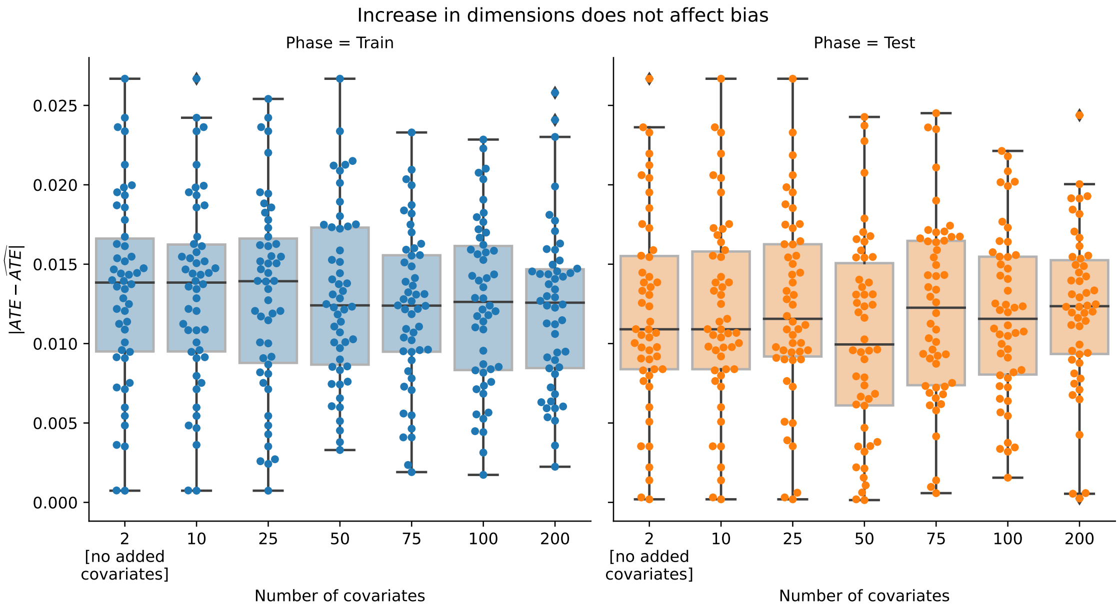


Figure C3: Comparison of estimation bias as a function of number of covariates. We take a synthetic dataset that adheres to conditional exchangeability and increasingly add noisy covariates to it. We see the estimation bias remains the same, throughout the train set (left) and the test set (right) regardless of the number of covariates added. This better strengthens our claim for the suitability of our method to discover unbiased natural experiments in high-dimensional data.

Partitioning The tree partitions so to recreate the four intended sub populations where a natural experiment was simulated. The average Rand index was equal to 0.901 across the 50 subsamples ($\sigma = 0.263$). Violating leaf nodes are marked in red.

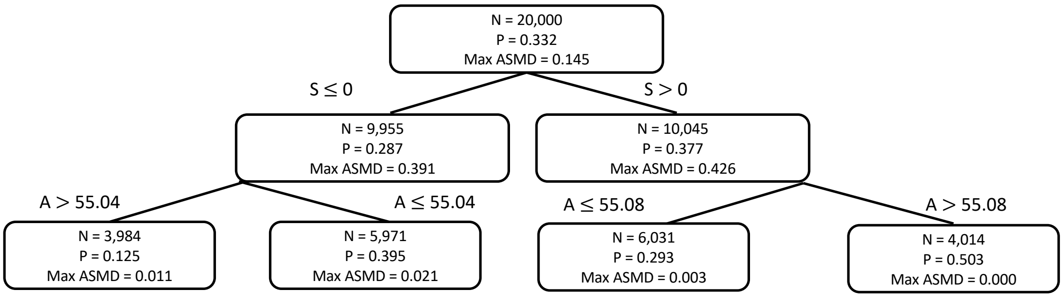


Figure C4: Tree structure after training on the entire natural experiment dataset ($N = 20,000$). Violating leaf nodes are marked in red. P is for node prevalence.

Causal effect estimation Here, Causal Tree has both higher estimation bias and higher estimation variance compared to other methods. The tree-like nature of our data structure may be incompatible with the optimization function of Causal Tree, which maximizes on treatment effect heterogeneity. Causal Forest, which has multiple Causal Tree models, however overcomes this difficulty and performs well.

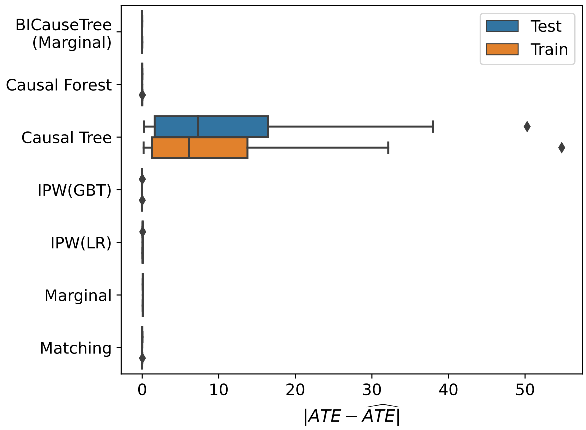


Figure C5: Estimation bias for the natural experiment dataset across 50 subsamples with $N = 20,000$.

Propensity score estimation BICauseTree(Marginal)'s propensity estimation is well calibrated. It is closer to the identity line than logistic regression- (IPW(LR)) and the gradient boosting trees- (IPW(GBT)) based models. The calibration however remains satisfactory across all models. This further shows our ability to identify natural experiments in a simple data setting.

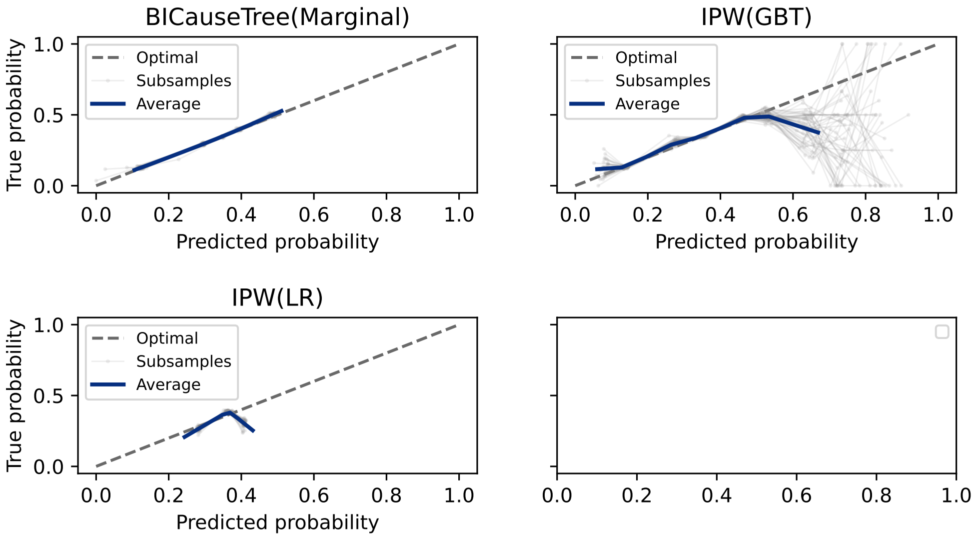


Figure C6: Calibration of the propensity score estimation for the natural experiment dataset across 50 subsamples, on the natural experiment testing set ($N = 10,000$).

Outcome estimation The performance of BICauseTree w.r.t outcome estimation is only satisfactory in this experiment. This shows our ability to estimate the ATE despite a somehow lower calibration of our outcome estimation. Ultimately, the performance would likely have been improved had we used a more complex outcome model. The calibration of the predicted outcomes by BICauseTree however remains better than the one by Matching.

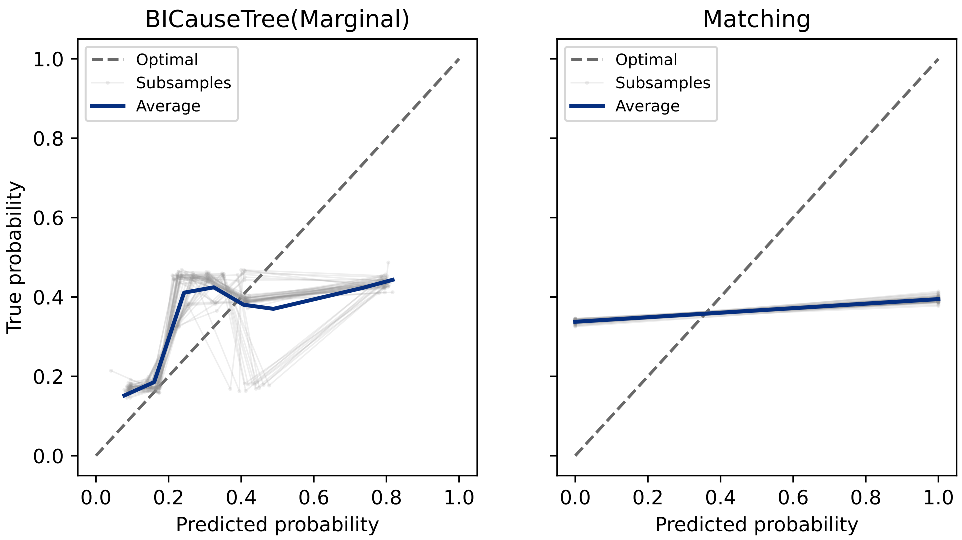


Figure C7: Calibration of the outcome estimation for the natural experiment dataset across 50 subsamples on the natural experiment testing set ($N = 10,000$).

Effect estimation bias reduction with tree depth Figure C8 shows how estimation bias decreases as we increase the maximum depth hyperparameter of our BICause-Tree(Marginal) on the natural experiment training dataset ($N = 10,000$). Here, each circle in the plot represents the effect estimated for each subsample. The dotted line shows the average bias with an IPW estimator. The shaded area represents the 95% confidence interval (CI) for IPW. This result shows that our estimation is robust to the choice of a maximum tree depth hyperparameter. Above a certain threshold, the effect estimate from BICauseTree remains unbiased.

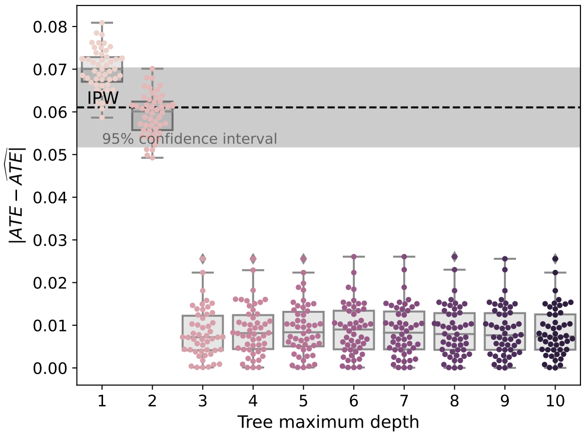


Figure C8: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the natural experiment training set ($N = 10,000$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure C9 below shows the weighted ASMD of both covariates S and A in BICauseTree models with varying maximum tree depth hyperparameters. It illustrates the reduction of treatment allocation bias as the tree grows, but also shows that this reduction is a heuristic as the ASMD might not decrease monotonically.

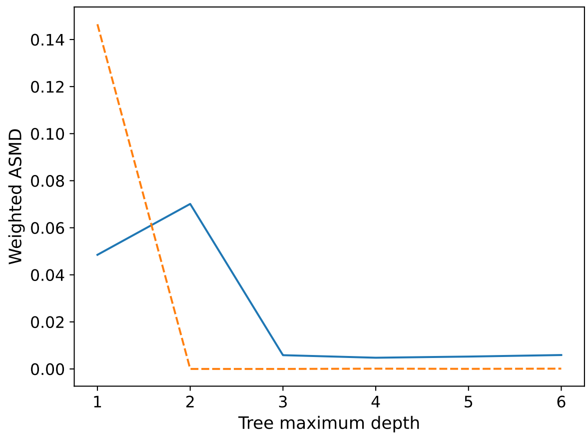


Figure C9: Weighted ASMD for all covariates applying *BICauseTree(Marginal)* models with varying maximum tree depths on the natural experiment training set ($N = 10,000$) across 50 subsamples.

C.5.2.2 Positivity violations dataset

Partitioning BICauseTree coherently flags the two subpopulations where a positivity violations were simulated, as shown in the example partition on the entire dataset below in Figure C10 (violating leaves are marked in red). The average Rand index was equal to 0.986 across the 50 subsamples ($\sigma = 0.026$).

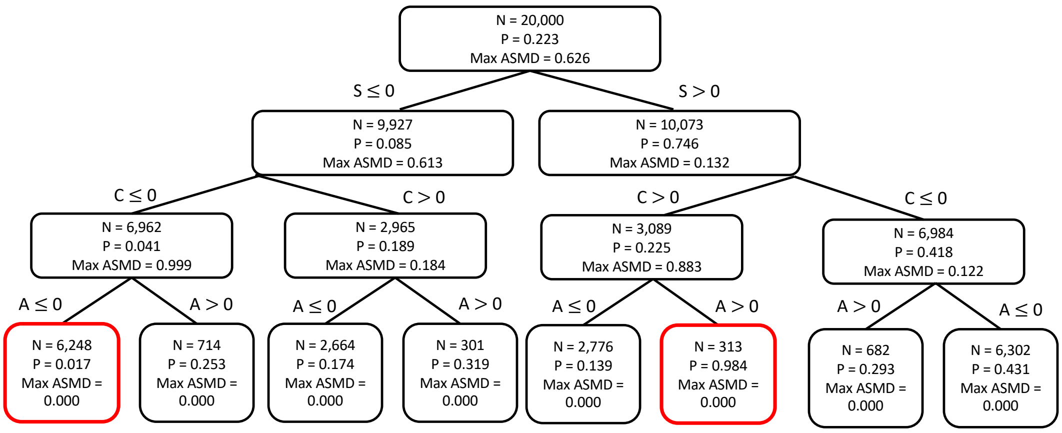


Figure C10: Tree structure after training on the entire positivity violations dataset ($N = 20,000$). Violating leaf nodes are marked in red.

Causal effect estimation Our tree compares with all existing alternatives. Matching has both higher estimation bias and higher estimation variance compared to other methods.

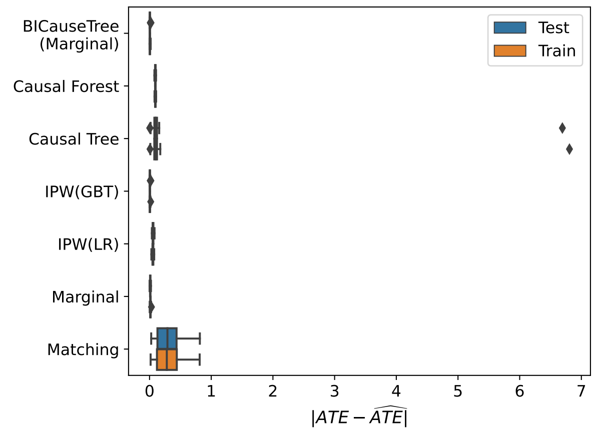


Figure C11: Estimation bias on the positivity violations dataset ($N = 20,000$) across 50 subsamples

Propensity score estimation BICauseTree(Marginal) shows good calibration, since it indeed clearly identified the correct subpopulations. The gradient-boosting tree (IPW(GBT)) also shows good calibration, as the tree-based modelling captures the underlying structure of the data. On the other hand, the logistic regression-based model (IPW(LR)) is ill-specified for the data and therefore presents relatively poorer calibration.

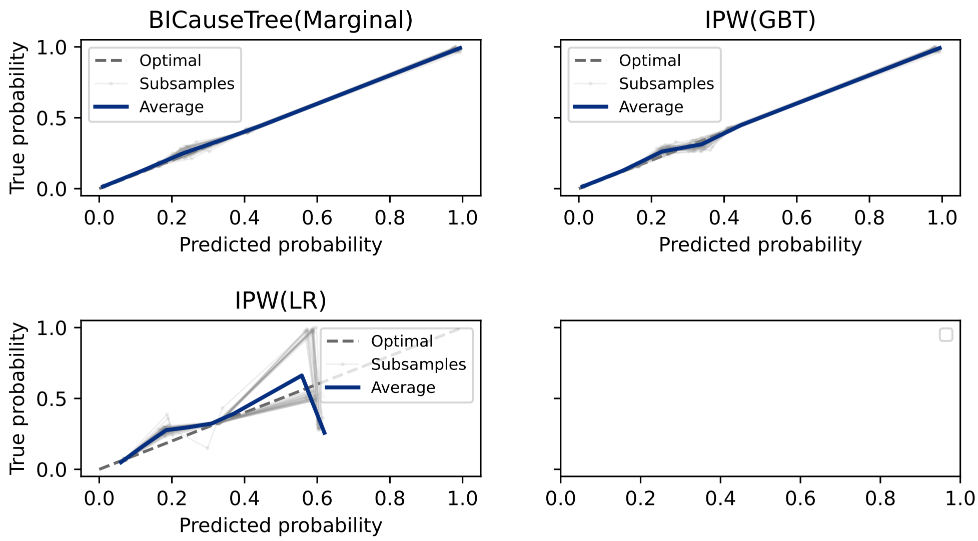


Figure C12: Propensity score calibration comparing our approach to existing alternatives on the testing set of the positivity violation dataset ($N = 10,000$) across 50 subsamples

Outcome estimation BICauseTree shows descent calibration on outcome prediction in this experiment. Matching shows poor calibration.

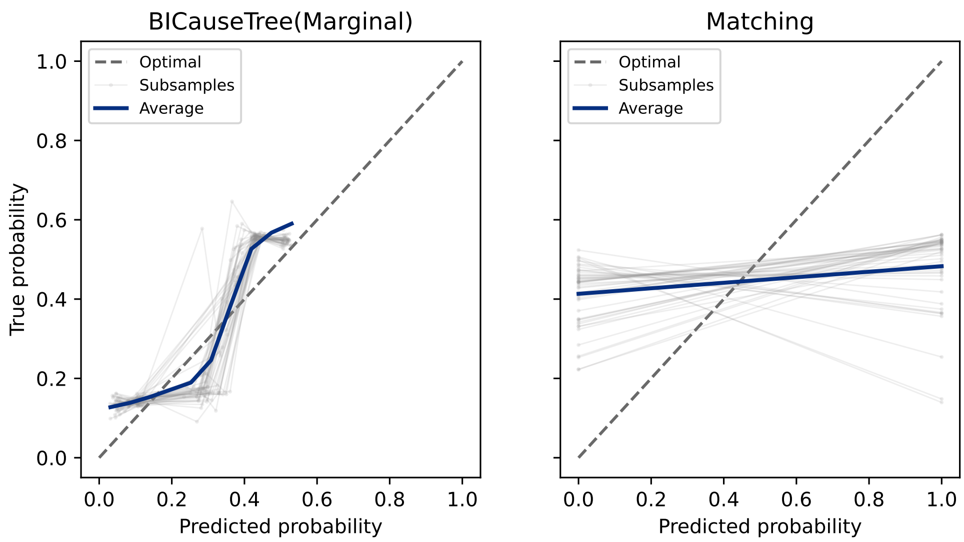


Figure C13: Calibration of the outcome estimation comparing our approach to existing alternatives on the positivity violations testing dataset ($N = 10,000$) across 50 subsamples

Effect estimation bias reduction with tree depth We see in this case, that having a max depth which is too large leads to an increase in the bias. This might indicate that a hyper parameter tuning procedure would benefit cases where the tree might become too deep and overfit to the training data.

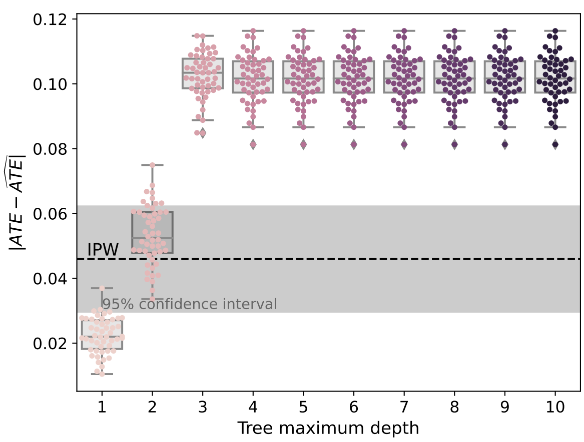


Figure C14: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the positivity violations experiment training set ($N = 10,000$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure C15 below shows the weighted ASMD of all three covariates in BICauseTree models with varying maximum tree depth hyperparameters. We see that the ASMD generally decreases with tree depth, showing that the splitting criteria we use leads to balanced subpopulations, as expected.

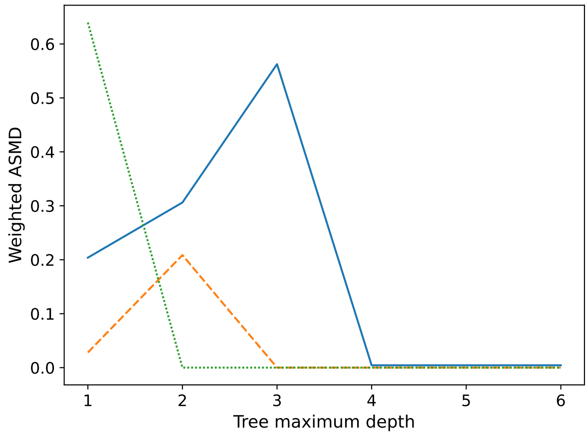


Figure C15: Weighted ASMD for all covariates applying *BICauseTree(Marginal)* models with varying maximum tree depths on the positivity violations dataset training set ($N = 10,000$) across 50 subsamples

C.6 Further experiment results: the twins dataset

Causal effect estimation Figure C16 shows a comprehensive comparison of the different models. Causal tree shows poor performance, whereas causal forest performs comparably to BICauseTree(Marginal) and IPW.

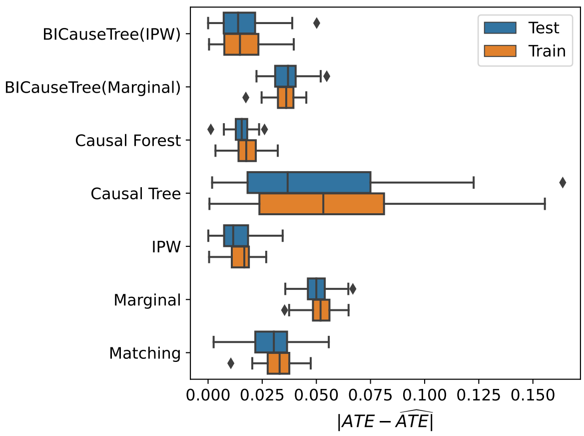


Figure C16: Estimation bias on the twins dataset ($N = 11,984$) across 50 subsamples

Partitioning The following plot C17 shows the final tree built for the twins dataset. Red nodes indicate positivity violation population.

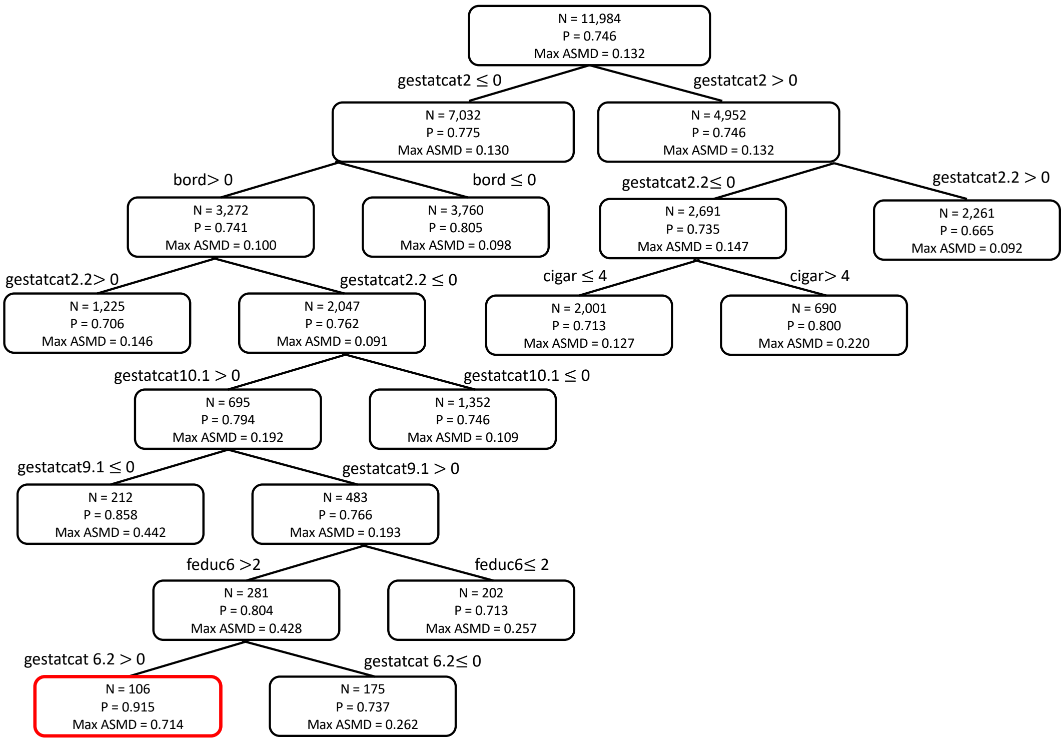


Figure C17: Tree structure after training on the entire twins dataset ($N = 11,984$). Violating leaf nodes are marked in red. P is for node prevalence. P is for node prevalence.

Outcome estimation BICauseTree(Marginal) shows good calibration on outcome estimation. BICauseTree(IPW) seem to have some bias in outcome estimation. This is possibly due to the parametric nature of our IPW model which uses a Logistic Regression. Matching shows poor outcome estimation ability.

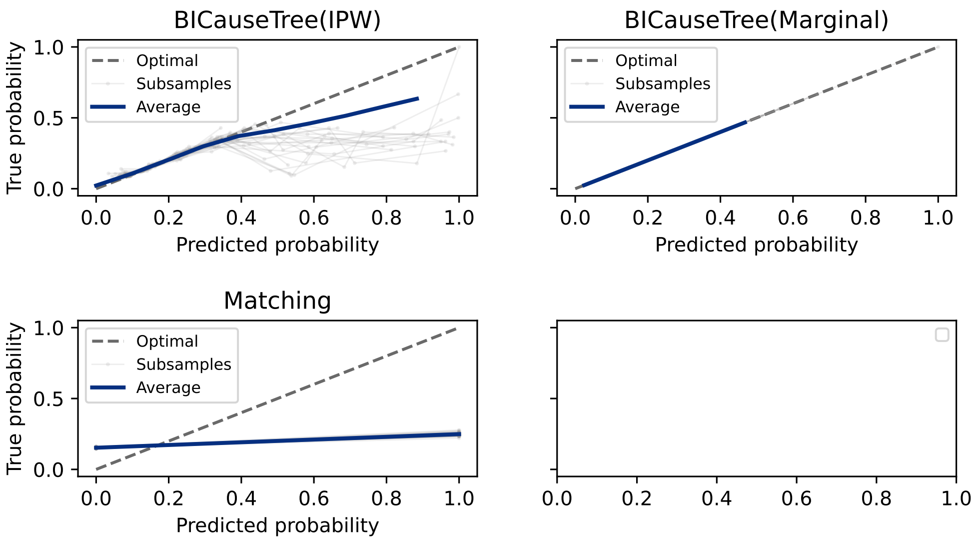


Figure C18: Calibration of the outcome estimation across 50 subsamples, on the twins testing set ($N = 5,992$).

Treatment allocation bias reduction with tree depth Figure C19 below shows the reduction in ASMD for the top 10 most imbalanced features in the entire population. It compares BICauseTree models with varying maximum tree depth hyperparameters. In all 10 covariates, we notice how the ASMD reduces as the maximum tree depth increases.

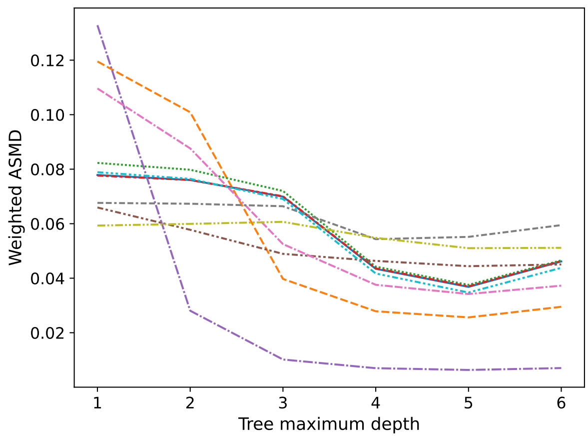


Figure C19: Weighted ASMD for the top 10 covariates with the highest ASMD in the initial population for *BICauseTree(Marginal)* models with varying maximum tree depths on the twins dataset training set ($N = 5,992$) across 50 subsamples

C.7 Further experiment results: the ACIC dataset

Causal effect estimation Figure C16 shows a comprehensive comparison of the different models. The performance of both BICauseTree models compares with the one from IPW or Causal Forest. Causal Tree however shows poor performance. Matching shows high estimation bias compared to all other models, including the Marginal estimator.

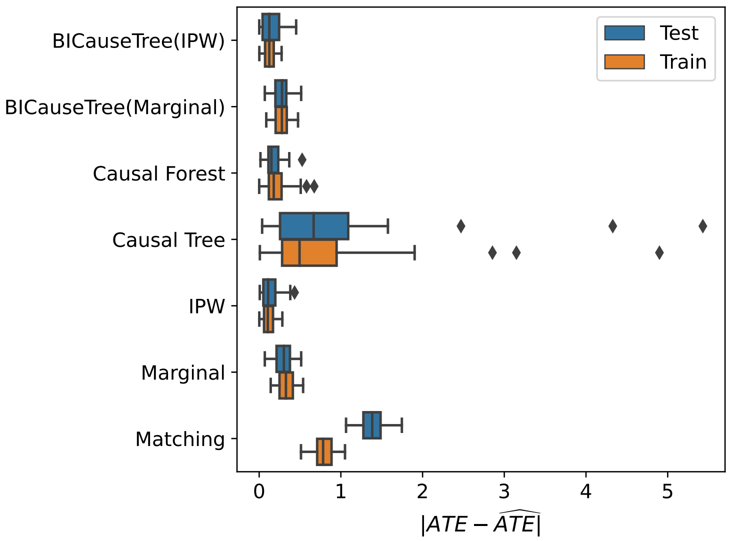


Figure C20: Estimation bias on the ACIC dataset testing set ($N = 960$) across 50 subsamples

Partitioning The following figure C22 shows the tree built for the ACIC dataset.

Propensity score estimation We show here the calibration of propensity scores in the ACIC dataset. BICauseTree preforms less well than IPW, which is more tailored for propensity estimation. As the outcome is non-binary in ACIC, we are not able to generate a calibration plot comparing the outcome estimation of BICauseTree with the one from existing approaches.

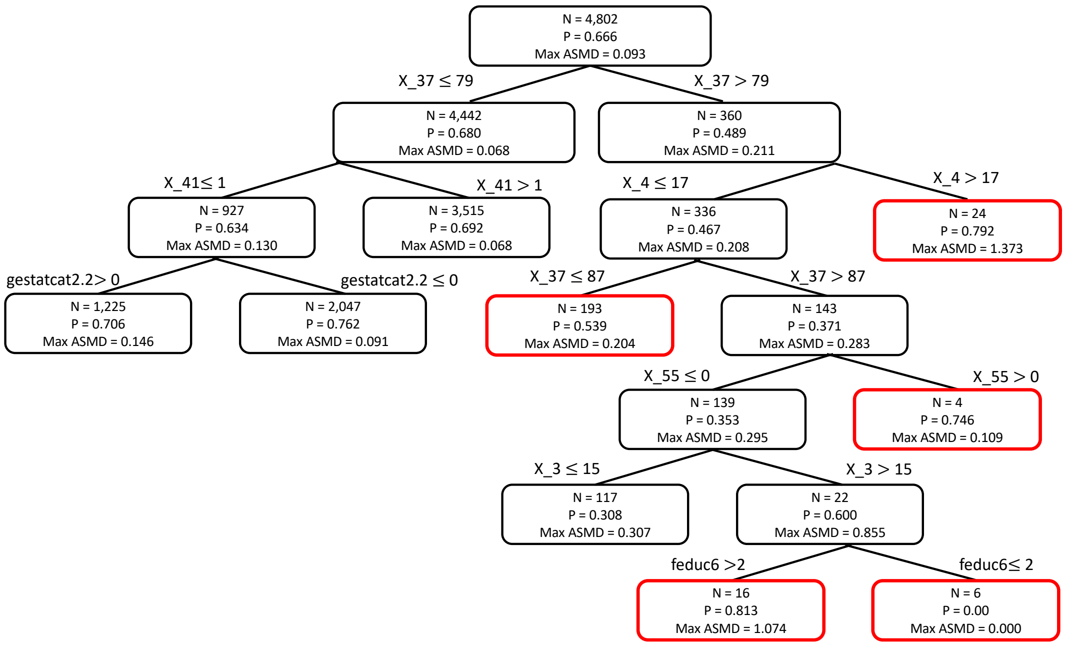


Figure C21: Tree structure after training on the entire ACIC dataset ($N = 4,802$). Violating leaf nodes are marked in red. P is for node prevalence. P is for node prevalence.

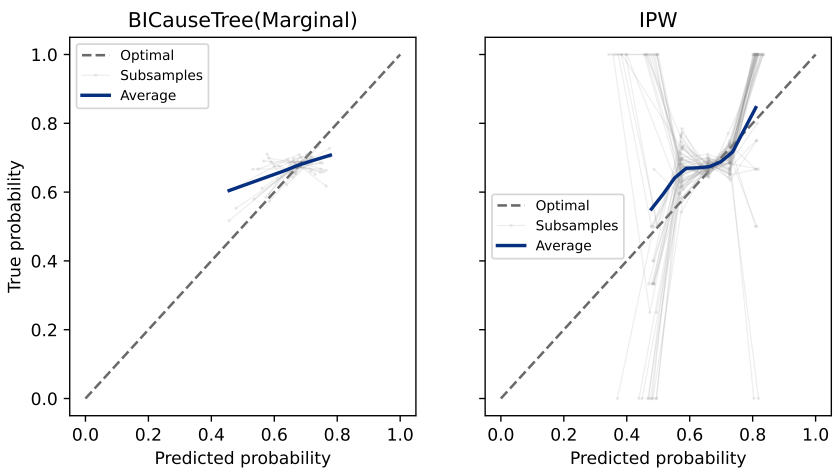


Figure C22: Propensity score calibration comparing our approach to existing alternatives on the ACIC dataset testing set ($N = 960$) across 50 subsamples

Effect estimation bias reduction with tree depth Figure C23 below shows the estimation bias for various maximum tree depth hyperparameters. We notice some overlap to IPW confidence interval, and bias reduces for depths up to max depth 6, then the variance across subsamples starts to increase. This may be due to the limited sample size of this dataset.

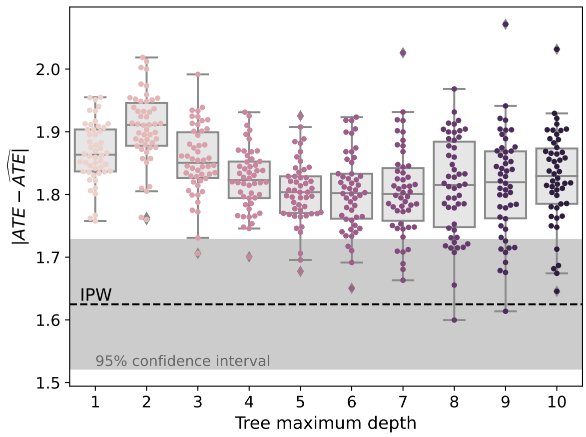


Figure C23: Estimation bias when comparing *BICauseTree(Marginal)* with varying maximum depth parameters with the average bias of IPW (dotted), on the ACIC training set ($N = 3,842$) across 50 subsamples.

Treatment allocation bias reduction with tree depth Figure C24 below shows the reduction in ASMD for the top 10 most imbalanced features in the entire population. It compares BICauseTree models with varying maximum tree depth hyperparameters. All 10 covariates show reduction in ASMD with depth.

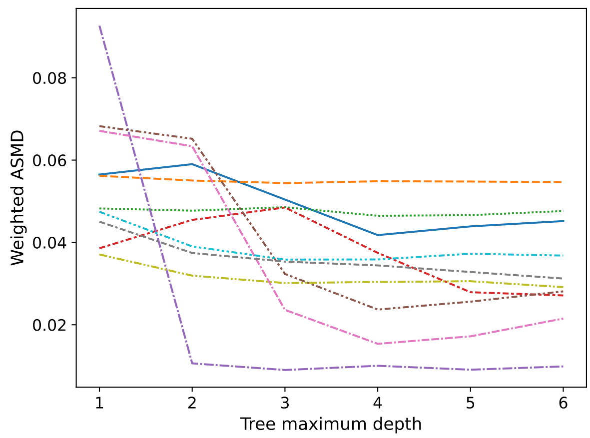


Figure C24: Weighted ASMD across maximum tree depths for the top 10 covariates with the highest ASMD in the initial population, applying *BICauseTree(Marginal)* models with varying maximum tree depths on the ACIC dataset training set ($N = 3,842$) across 50 subsamples

C.8 User study

We interviewed 25 participants had expertise in Statistics, but no expertise in causal inference. Non-binary responses were given on a Likert scale from 1 (Very unclear) to 5

(Very clear). All participants roughly knew about causal inference, and had a clear brief on the methods compared.

1. **Clarity**
   Question: How clear was the explanation of the average treatment effect provided by this method?
   Follow-up: Please describe any part of the explanation you found unclear.

2. **Trustworthiness**
   Question: How trustworthy do you find the estimated average treatment effect based on the explanation provided?
   Follow-up: What, if anything, made you question the trustworthiness of the explanation?

3. **Actionability**
   Question: Based on the explanation, how confident are you that you could make a decision or policy change involving the treatment effect?
   Follow-up: What additional information, if any, would improve your ability to act on this explanation?

4. **Causal Understanding**
   Task-based Question: Summarize the relationship between the treatment and the outcome as described in the explanation. What is the estimated average treatment effect?
   Scoring: Does the treatment increase or decrease the outcome, and by how much?

5. **Low cognitive Load**
   Question: How easy was it to understand the explanation of the treatment effect?
   Follow-up: Which part of the explanation, if any, did you find most challenging?

6. **Time to Understand**
   Task-based Measurement: Track the time participants take to read and summarize the explanation.
   Question: Do you feel you had enough time to fully understand the explanation?
   Scale: Yes/No
   Follow-up: How much more time would you need to feel confident in your understanding?

7. **Intuitiveness**
   Question: How intuitive did you find the method used to compute the average treatment effect in this explanation?

Follow-up: What aspects of the method made it more or less intuitive to understand the treatment effect estimation?

Method Comparison Table

Section	Matching	IPW	Causal Tree	BICause Tree
Clarity	3.84	3.08	2.84	4.68
Trustworthiness	3.88	4.28	4.0	3.32
Actionability	3.76	4.04	3.36	3.6
Causal Understanding	4. 76	4.8	3.88	4.72
Low cognitive Load	4.64	3.76	2.48	4.64
Time to Understand	1 min	2 min	3 min	1 min
Intuitiveness	4.08	3.72	2.92	4.8

Table C2: Comparison of interpretability metrics across four methods, on a Likert scale from Not very likely (1) to very likely (5). For Time to Understand, we report the most common response. Averaged across $N = 25$ participants.

C.9 Implementation

C.9.1 Positivity violations evaluations

We implemented two positivity violations definition procedures: the Crump method and a method we introduced, which we call the *symmetric prevalence* threshold method. The Crump method is a data-driven method that defines a threshold for extreme propensity scores (i.e., positivity violations), based on the distribution of propensity scores in the data (see further results in [37]).

The symmetric prevalence procedure generates upper and lower cutoff values that are adjusted for the prevalence. The cutoffs are computed such that if the overall prevalence was 0.5 they would be symmetrical (e.g. for $\alpha = 0.05$ the cutoffs would be 0.05 and 0.95). We may consider this as class reweighting of the propensities, within the entire population. If we denote the overall prevalence μ, and consider α as the cutoff had the distribution been symetric (we recommend taking $\alpha = 0.1$), the cutoffs are computed as follows:

$$\textit{Upper cutoff} = \frac{(1-\alpha)*\mu}{(1-\alpha)*\mu+\alpha*(1-\mu)}$$

$$\textit{Lower cutoff} = \frac{\alpha*\mu}{\alpha*\mu+(1-\alpha)*(1-\mu)}$$

C.9.2 Experimental details and computation

For BICauseTree, most hyperparameters were set to their default value. Multiple hypothesis test correction was done following a step-down method using Holm-Bonferroni adjustments [131, 132], with $\alpha = 0.05$. The threshold for weight trimming for positivity violations

was computed using the Crump procedure [37, 247] with 10000 segments. Throughout all experiments the minimum treatment group size was set to 2 patients. The maximum depth is the only parameter which varied depending on the experiments. It was set to 5 for both synthetic datasets and 10 for the experiments on the twins dataset. A long-standing practice has been to define any covariate with $ASMD \geq 0.10$ as a potentially problematic confounder [129], so we set this as our default threshold and used it in our experiments.
For IPW(LR), we used a Logistic Regression with a *saga* solver, no penalty and a maximum number of iterations equal to 500. For comparison purposes, BICauseTree(IPW) used similar hyperparameters for its internal IPW outcome model. For IPW(GBT) we used default hyperparameters.
We applied double matching based on the Mahalanobis distance for all datasets but ACIC, on which we used an Euclidean distance to avoid non-invertable matrices caused by the sparsity of non-null column values.
Causal Tree with a single estimator and a subforest size of 1. For Causal Forest, we used 50 estimators and a subforest size of 1.
For including Causal Tree into our experiments, we used the code available at `https://github.com/py-why/EconML`. Outcome and propensity models were trained using `sklearn` with default parameters and 500 maximum iterations for Logistic Regression when relevant. Statistical testing was implemented using the `statsmodel` package.

C.9.2.1 Compute and Runtime

In general, BICauseTree inherits its computational efficiency from decision trees. Namely the fitting complexity is $O(M \cdot V \cdot N\log(N))$, Where N it the number of observations, M is the number of features, and $V = \max_{m \in M}(V_m)$ the maximal number of unique levels among all M features. Briefly, recursively splitting the tree is $O(N\log(N))$, Additionally, for every recursive iteration we consider all M features (calculating ASMD is $O(N)$ for each feature), and, among the select feature we consider a split cutoff over all its unique values (note this is 1 for binary features, calculating a statistical test takes $O(N)$ for each level). Therefore the resulting complexity is $O(MVN\log(N))$.
Table C3 shows the average runtime it took for each model to fit. We see BICauseTree compares with IPW, Causal Tree and Causal Forest in terms of wall-clock run-time, while Matching has higher compute time than all other models.
We further tested the runtime of BICauseTree as a function of dimensionality. Figure C25 shows the time, averaged over 10 random splits, it took for BICauseTree to fit—including filtering for positivity—as a function of the number of covariates in the input data (following the experiment design presented earlier in the Appendix in Figure C3.

Experiment	Amount of Compute
Natural experiment dataset	
BICauseTree(Marginal)	286
IPW	134
Matching	882
Causal Tree	183
Causal Forest	390
Positivity violations dataset	
BICauseTree(Marginal)	258
IPW	128
Matching	929
Causal Tree	137
Causal Forest	384
Twins	
BICauseTree(Marginal)	420
BICauseTree(IPW)	567
IPW	329
Matching	2283
Causal Tree	403
Causal Forest	672
ACIC	
BICauseTree(Marginal)	329
BICauseTree(IPW)	376
IPW	239
Matching	1092
Causal Tree	354
Causal Forest	439

Table C3: Total amount of compute in seconds for model fitting across 50 train-test splits, for selected experiments. All experiments were run on a 10 core CPU Apple M1 Pro.

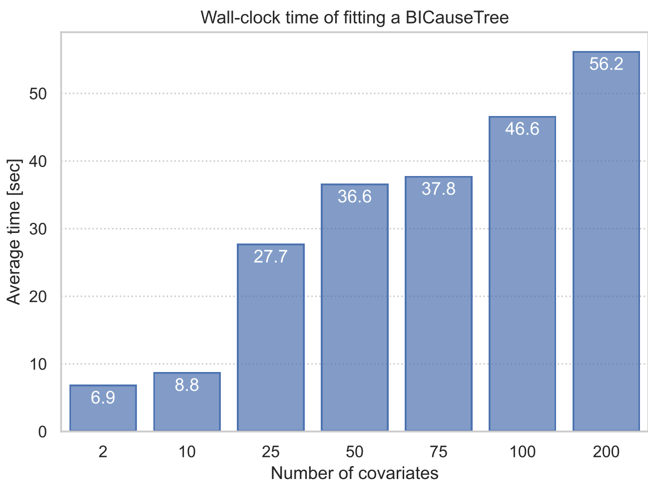


Figure C25: Runtime of BICauseTree as a function of increasing the dimensionality of the input. The bars show the average wall-clock time, in seconds, it took to fit a BICauseTree for different numbers of input covariates.

C.9.2.2 Causal benchmark datasets

The **twins dataset** was originally taken from the denominator file at `https://www.nber.org/research/data/linked-birthinfant-death-cohort-data`. However, we use data generated by *Neal et. al* [138], which simulates an observational study from the initial data by selectively hiding one of the twins with a generative approach. The sample size is $N = 11,984$ pairs of twins, with the essential inclusion criterion being that both individuals were born weighing less than 2kg. The mortality rate amongst the lighter twins is 18.9%, and for the heavier 16.4%, for an average treatment effect of -2.5% (which we thus consider as ground truth). A total of 75 covariates were recorded, relating to the parents' socio-demographic features and medical history, the pregnancy and the birth.
The **ACIC dataset** was generated using the *causallib* package at `https://github.com/BiomedSciAI/causallib/blob/master/causallib/datasets/data/acic_challenge_2016/README.md`. It contains covariates, simulated treatment, and simulated response variables for the causal inference challenge in the 2016 Atlantic Causal Inference Conference [248]. For each of 20 conditions, treatment and response data were simulated from real-world data corresponding to 4802 individuals and 58 covariates. After one-hot encoding, a total of 79 covariates was included. More specifically, we used the set of treatment and response variables *zymu 174570858*, with the two expected potential outcomes (*mu0*, *mu1*).

C.10 Social impact of our work: further details

Recent years have seen a surge in the Explainable AI (XAI) literature, motivated by rising ethical concerns around artificial intelligence. Model scrutiny is particularly relevant in sensitive environments such as healthcare, which require high safety standards considering the major consequences of invalid predictions on individual trajectories [249]. Calls for model transparency are further motivated by evidences that supervised machine learning is inclined to reproduce inherent bias and prejudice against discriminated groups [250]. Ultimately, XAI aims at building trust in models that ought to be deployed. In recent years, the demand for transparency expanded beyond the research community, notably with incentives from high institutions. The European Union General Data Protection Regulation legislation has mandated a "right to explanation" for individual predictions that can "significantly affect" users [12]. In other words, algorithmic results should be re-traceable on demand. In parallel, quality control frameworks such as the U.S. Food and Drug Administration guidance have been introduced to ensure the safety of clinical AI [13]. By providing interpretable effect estimation, our BICauseTree approach aligns with the mission of increasing transparency for downstream users. As such, it is most likely to have positive social impact. We however caution against relying exclusively on causal inference when data accuracy or volumes are insufficient, as misleading effect estimates may have negative social impact.

Appendix D

Supplementary Materials: Is merging worth it ? Securely evaluating the information gain for causal dataset acquisition

D.1 Mathematical Details

D.1.1 Algorithms for Computing EIG

Algorithm 3 Algorithm for $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e)$

Input: Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2

Output: $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e)$

Require: $l = NM_1$ for some $n, M_1 > 0$

$S \leftarrow 0$

Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathscr{D}_0)$ for $l = NM_1$

Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i), (\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$

for $i \in \{1, \cdots, N\}$ **do**

  Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e, \mathbf{t}_e, \theta_i)$

  $S \leftarrow S + \log \frac{P(\mathbf{y}_e^{(i)}|\theta_i, \mathbf{X}_e, \mathbf{t}_e)}{\frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j}, \mathbf{X}_e, \mathbf{t}_e)}$

$\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e) \leftarrow \frac{1}{N} S$

return $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e)$

Algorithm 4 Algorithm for $\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathscr{D}_0}(e)$

Input:$\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathscr{D}_0)$, Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathscr{D}_0}(e)$

$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{l} \sim P(\theta \mid \mathscr{D}_0)$ for $l = NM_1$
Split $\{\theta_i\}_{i=1}^{l}$ as $\{(\theta_i),(\theta_{i,j})_{j=1}^{M_1}\}_{i=1}^{N}$
for $i \in \{1,\cdots,N\}$ **do**
  Sample $\mathbf{y}_e^{(i)} \sim P(\mathbf{y}_e|\mathbf{X}_e,\mathbf{t}_e,\theta_i)$
  $S \leftarrow S - \log \frac{1}{M_1}\sum_{j=1}^{M_1} P(\mathbf{y}_e^{(i)}|\theta_{i,j},\mathbf{X}_e,\mathbf{t}_e)$
$\widehat{\mathrm{EIG}}^{\mathrm{RB}}_{\theta|\mathscr{D}_0}(e) \leftarrow \frac{1}{N}S$
return $\widehat{\mathrm{EIG}}^{\mathrm{NMC}}_{\theta|\mathscr{D}_0}(e)$

Algorithm 5 Algorithm for $\widehat{\mathrm{EIG}}_{\theta_c|\mathscr{D}_0}(e)$

Input Data Matrix $\mathbf{X}_e$, Treatment Vector $\mathbf{t}_e$, Variance σ^2
Output: $\widehat{\mathrm{EIG}}_{\theta_c|\mathscr{D}_0}(e)$

$S \leftarrow 0$
Sample $\{\theta_i\}_{i=1}^{N} \sim P(\theta \mid \mathscr{D}_0)$
for $i \in \{1,\cdots,N\}$ **do**
  Sample $\mathbf{y}_e \sim P(\mathbf{y}_e|\theta,\mathbf{X}_e,\mathbf{t}_e)$
  Sample $\{\theta'_j\}_{j=1}^{M_1} \sim P(\theta \mid \mathscr{D}_0)$
  Sample $\{(\theta_{\mathrm{nc}})^{(ik)}\}_{k=1}^{M_2} \sim P(\theta_{\mathrm{nc}} \mid (\theta_c)_i,\mathscr{D}_0)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0) \leftarrow \frac{1}{M_1}\sum_{j=1}^{M_1} P\left(\mathbf{y}_e^{(i)}|\theta'_j,\mathbf{X}_e,\mathbf{t}_e\right)$
  $\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)},\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0) \leftarrow \frac{1}{M_2}\sum_{k=1}^{M_2} P\left(\mathbf{y}_e^{(i)}|\theta_{nc}^{(ik)} \cup \theta_c^{(i)},\mathbf{X}_e,\mathbf{t}_e\right)$
  $S \leftarrow S + \log\left(\frac{\widehat{P}(\mathbf{y}_e^{(i)}|\theta_{\mathrm{c}}^{(i)},\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}{\widehat{P}(\mathbf{y}_e^{(i)}|\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}\right)$
return $\frac{1}{N}S$

D.1.2 Differential Privacy Definition

Definition 6. We say a randomised algorithm, $\mathscr{A}$ satisfies ε differential privacy if for any input dataset $\mathscr{D}$ and dataset $\mathscr{D}'$ differing by a single entry, we have

$$P(\mathscr{A}(\mathscr{D}) \in \mathscr{O}) \leq \exp(\varepsilon)P(\mathscr{A}(\mathscr{D}') \in \mathscr{O}).$$

D.1.3 Sensitivity of Linear Statistic

We derive the sensitivity of the linear EIG statistic in order to give a fair comparison with naive differential privacy in Section 6.5.3.

Proposition 5. *Let:*

$$f(\mathbf{X}_e) = \log\det(\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0) \tag{D.1}$$

If $\Lambda_0 = cI$ *and* $\|\mathbf{x}\|_\infty \leq M$ *for all* $\mathbf{x} \sim P_e(\mathbf{x})$ *and e. Then we have that:*

$$\Delta_f = \max_{\mathbf{X}'_e, \mathbf{X}_e} \|f(\mathbf{X}'_e) - f(\mathbf{X}_e)\| \leq \frac{Md}{\sqrt{c}} \tag{D.2}$$

Where $\mathbf{X}'_e, \mathbf{X}_e$ *differ in at most one row. This implies* $f(\mathbf{X}_e) + Z$ *for* $Z \sim \mathrm{Laplace}(\frac{Md}{\varepsilon\sqrt{c}})$ *is a* ε *differentially private release of* $f(\mathbf{X}_e)$*.*

Proof. To prove this we will use the fact that if $\max_{\|\mathbf{X}\| \leq M} \|Df(\mathbf{X})\|_F = L$ we have that $\|f(\mathbf{X}'_e) - f(\mathbf{X}_e)\| \leq L\|\mathbf{X}'_e - \mathbf{X}_e\|_F$ for all $\mathbf{X}'_e, \mathbf{X}_e$ with norm bounded by M where $\|\cdot\|_F$ is the Frobenious norm. Write $\mathbf{E} = \mathbf{X}_0^\top \mathbf{X}_0 + \Lambda_0$, via the chain rule we have that $Df(\mathbf{X}) = \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E} \rangle^{-1} \mathbf{X}_e^\top$ Now:

$$\|\langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1} \mathbf{X}_e^\top\|_F = \sqrt{\mathrm{tr}\langle \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1} \mathbf{X}_e^\top \mathbf{X}_e \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1}\rangle} \tag{D.3}$$

$$= \sqrt{\mathrm{tr}\langle \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1} - \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1} \mathbf{E} \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1}\rangle} \tag{D.4}$$

$$\leq \sqrt{\mathrm{tr}\langle \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1}\rangle} \leq \sqrt{\frac{d}{c}} \tag{D.5}$$

We have used the fact that $\langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1} \mathbf{E} \langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1}$ is positive semi definite and so has positive trace, and that the eigenvalues of $\mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}$ are bounded below by c so the eigenvalues of $\langle \mathbf{X}_e^\top \mathbf{X}_e + \mathbf{E}\rangle^{-1}$ are bounded above by $\frac{1}{c}$. Finally for neighbouring datasets we can change at most d entries by M so $\|\mathbf{X}'_e - \mathbf{X}_e\|_F \leq \sqrt{d}M$ □

D.2 Model Details

Throughout all methods will aim to model the outcome as some function plus an error, so:

$$Y_i = f(\mathbf{x}_i, t_i) + \varepsilon_i \tag{D.6}$$

D.2.1 Models via Sampling

Bayesian Additive Regression Trees (BART) BART models [178] f as the sum of L piecewise constant binary regression trees, so we have:

$$f(\mathbf{x}, t) = \sum_{l=1}^{L} g_i(\mathbf{x}, t, T_l, \mathbf{m}_l) \tag{D.7}$$

where T_l is a regression tree given by a partition $(\mathscr{A}_1, \cdots, \mathscr{A}_{\mathscr{B}(l)})$ of $\mathscr{X} \times \mathscr{T}$ and the set of leaf parameter values $\mathbf{m}_l = (m_{l1}, \cdots, m_{l\mathscr{B}(l)})$ so that:

$$g_l(\mathbf{x}, t) = m_j \text{ if } \mathbf{x}, t \in \mathscr{A}_j \tag{D.8}$$

The mean parameters are given with independent normal parameters $m_{lj} \sim \mathscr{N}(0, \sigma_m)$. Over trees, the prior is such that the probability of a node having children at depth d is given by:

$$\alpha(1+d)^{-\theta} \text{ for } \alpha \in (0,1), \theta \in [0, \infty) \tag{D.9}$$

The original BART model explores this space using Metropolis-Hastings Markov chain Monte Carlo, but we make use of XBART [251] for accelerated posterior sampling.

Bayesian Causal Forest Bayesian Causal Forests [BCF; 163] build upon BART models utilising specific parameterisations for causal inference tasks. The two parameterisations suggested in Hahn et al. [163] are firstly:

$$f(\mathbf{x}, t) = \mu(\mathbf{x}) + t\tau(\mathbf{x}) \tag{D.10}$$

Where μ, τ are independent BART models. To draw specific attention to their parameters will write them as $\mu_{\theta_{\text{nc}}}, \tau_{\theta_{\text{c}}}$ noting that $\tau_{\theta_{\text{c}}}$ is a model for CATE. Hahn et al. [163] note that this parameterisation is not invariant to which treatment is assigned as positive or negative, leading them to propose the following invariant parameterisation:

$$f_\theta(\mathbf{x}, t) = \tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{D.11}$$

Where $b_t \sim \mathscr{N}(0, \frac{1}{2})$. Under this parameterisation a CATE estimate is given by $(b_1 - b_0)\tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x})$. When sampling we make use of the accelerated BCF approach [252] which builds upon XBART and uses the following slightly modified model:

$$f_\theta(\mathbf{x}, t) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) + b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) \tag{D.12}$$

Where $a \sim \mathscr{N}(0, 1)$.

We define the set θ_c to be any parameters affiliated with the τ model, including b_t for the invariant parameterisation. In order to sample $P(\theta_{\text{nc}}|\theta_c, \mathscr{D}_0)$ we refit θ_{nc} parameters on the dataset $\mathscr{D}_0$ to the residuals resulting from subtracting the τ portion of the model. So this refitting μ is as follows for the standard parametisation:

$$Y - t\tau_{\theta_{\text{c}}}(\mathbf{x}) = \mu_{\theta_{\text{nc}}}(\mathbf{x}) \tag{D.13}$$

Or for the accelerated BCF approach [252]:

$$Y - b_t \tilde{\tau}_{\tilde{\theta}_{\text{c}}}(\mathbf{x}) = a\tilde{\mu}_{\tilde{\theta}_{\text{nc}}}(\mathbf{x}) \tag{D.14}$$

Where any parameters on the left hand side are fixed.

D.2.2 Closed Form Models

In this section we give details of models for which the EIG is available in closed form. We provide details of the models as well as proofs for the expressions.

D.2.2.1 Bayesian Polynomial Regression Derivations

In this we derive the results for Bayesian polynomial regression. We have modelled our data as:

$$y \sim \mathcal{N}(\phi(\mathbf{x},t)^\top \theta, \sigma^2), \quad \theta \sim \mathcal{N}(\mu_0, \sigma^2 \Lambda_0^{-1}) \tag{D.15}$$

In this context, the posterior is available in closed form as:

$$\theta | \mathscr{D}_0 \sim \mathcal{N}(\mu_0, \sigma^2 \tilde{\Lambda}_0^{-1}) \tag{D.16}$$

$$\tilde{\Lambda}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right) + \Lambda_0^{-1}\right) \tag{D.17}$$

$$\mu_0 = (\Lambda_0)^{-1} \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right) \hat{\theta}_0 + \Lambda_0 \mu_0\right) \tag{D.18}$$

$$\hat{\theta}_0 = \left(\phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)\right)^{-1} \phi\left(\mathbf{X}_0, \mathbf{t}_0\right)^\top \mathbf{y}_0 \tag{D.19}$$

Expected Information Gain Over all parameters We use the fact that $\mathrm{EIG}_{\theta|\mathscr{D}_0}$ can be written as:

$$\mathrm{EIG}_{\theta|\mathscr{D}_0}(e) = \mathbb{E}_{P(\mathbf{y}_e|\mathbf{X}_e,\mathbf{t}_e,\mathscr{D}_0)}\left[H[P(\theta|\mathscr{D}_0)] - H[P(\theta|\mathscr{D}_0,\mathscr{D}_e)]\right],$$

As the posterior over θ is Gaussian we can directly evaluate these expressions using the closed form entropy for a Gaussian distribution as:

$$H[P(\theta|\mathscr{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\tilde{\Lambda}_0^{-1}\right)\right)$$

The distribution $\theta|\mathscr{D}_0, \mathscr{D}_e)$ can be obtained as above, where we now use $\tilde{\Lambda}_0$ as the prior precision matrix before updating on $\mathscr{D}_0$. This gives:

$$H[P(\theta|\mathscr{D}_0)] = \frac{n_e}{2}\left(1 + \log(2\pi)\right) + \frac{1}{2}\left(\log\det\left(\left(\phi\left(\mathbf{X}_e, \mathbf{t}_e\right)^\top \phi\left(\mathbf{X}_e, \mathbf{t}_e\right) + \tilde{\Lambda}_0\right)^{-1}\right)\right)$$

Using the above form for $\tilde{\Lambda}_0$ and collecting all constants gives the expression presented in the text.

$\mathrm{EIG}_{\theta_c|\mathscr{D}_0}$**: Expected Information Gain Over all parameters** This follows as above but using the fact that the block form of the matrices to allow us write the covariance precision matrix for the post host posterior over θ_c as follows:

$$(\mathbf{t}_0 \odot \phi_{\mathrm{c}}(\mathbf{X}_0))^\top (\mathbf{t}_0 \odot \phi_{\mathrm{c}}(\mathbf{X}_0)) + (\Lambda_0)_{[\mathrm{c,c}]}$$

Where $(\Lambda_0)_{[\mathrm{c,c}]}$ corresponds to block of Λ_0 with entries after $[c,c]$ in the row and column.

D.2.3 Causal Multitask Gaussian Processes [1]

In this work, CATE is modelled using a multitask Gaussian process [253]. Multitask Gaussian Processes use a GP in vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) to share information between tasks [180]. In Alaa and Van Der Schaar [1], learning the conditional outcome function for each treatment is seen as a separate task, so we jointly model:

$$Y|\mathbf{x},t \sim \mathcal{N}(0, f_t(\mathbf{x}), \sigma_t^2) \tag{D.20}$$

Where each f_t is a Gaussian Process. The kernel $\mathbf{K}_\eta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$ is now a symmetric positive semi-definite matrix-valued function, with hyper-parameters η. In the case of Alaa and Van Der Schaar [1] they use a *linear model of coregionalization*[1], giving the kernel as:

$$\mathbf{K}_\eta(\mathbf{x},\mathbf{x}') = \mathbf{A}_0 k_0(\mathbf{x},\mathbf{x}') + \mathbf{A}_1 k_1(\mathbf{x},\mathbf{x}') \tag{D.21}$$

Where k_t is the RBF kernel, given by:

$$k_t(\mathbf{x},\mathbf{x}') = \exp(-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}'\right)^\top \mathbf{R}_t^{-1}\left(\mathbf{x}-\mathbf{x}'\right)) \tag{D.22}$$

Where $\mathbf{R}_t^{-1} = \operatorname{diag}\left(\ell_{1,t}^2, \cdots, \ell_{d,t}^2\right)$ and $\ell_{j,t}$ is the length-scale parameter for the treatment $T = t$ in the j^{th} coordinate of $\mathbf{x}$. $\mathbf{A}_t$ is given by:

$$\mathbf{A}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{D.23}$$

Where $\theta_i j$ and ρ_i determine the variances and covariances of the shared tasks f_t. So we have that the full set of hyper-parameters $\eta = (\theta_0, \theta_1, \mathbf{R}_0, \mathbf{R}_1, \mathbf{A}_0, \mathbf{A}_1)$. Once all these hyper-parameters have been learnt we have that the covariance between different function evaluations, $f_t(\mathbf{x}), f_{t'}(\mathbf{x}')$, is given the t,t' coordinate of $\mathbf{K}_\eta(\mathbf{x},\mathbf{x}')$. So:

$$\operatorname{cov}(f_t(\mathbf{x}), f_{t'}(\mathbf{x}')) = \mathbf{K}_\eta(\mathbf{x},\mathbf{x}')_{[t,t']} \tag{D.24}$$

Now, if we let $K((\mathbf{x},t),(\mathbf{x}',t') = \mathbf{K}_\eta(\mathbf{x},\mathbf{x}')_{[t,t']}$ then we can obtain the posterior kernel in a similar way to the standard case. Precisely if we the training data be given by:

$$\tilde{\mathbf{X}} = \left[\{\mathbf{x}_i\}_{T_i=0}, \{\mathbf{x}_i\}_{T_i=1}\right]^T, \tag{D.25}$$

$$\tilde{\mathbf{Y}} = \left[\left\{y_i^{(T_i)}\right\}_{T_i=0}, \left\{y_i^{(T_i)}\right\}_{t_i=1}\right]^T, \tag{D.26}$$

$$\boldsymbol{\Sigma} = \operatorname{diag}\left(\sigma_0^2 \mathbf{I}_{n-n_1}, \sigma_1^2 \mathbf{I}_{n_1}\right) \tag{D.27}$$

[1]See Alvarez et al. [180] for more details.

$$n_1 = \sum_i W_i, \tag{D.28}$$

$$\mathbf{K}_\eta(x) = (\mathbf{K}_\eta\,(x, X_i)_i\,.) \tag{D.29}$$

Then we have that the posterior mutlitask GP has mean and posterior kernel given by:

$$m^{\text{post}}(\mathbf{x}) = \mathbf{K}_\eta^T(\mathbf{x})\,(\mathbf{K}_\eta(\mathbf{X},\mathbf{X}) + \Sigma)^{-1}\,\tilde{\mathbf{Y}} \tag{D.30}$$

$$\mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x},\mathbf{x}') = \mathbf{K}_{\eta^*}(\mathbf{x},\mathbf{x}') - \mathbf{K}_{\eta^*}(\mathbf{x})\,(\mathbf{K}_\eta(\mathbf{X},\mathbf{X}) + \Sigma)^{-1}\,\mathbf{K}_{\eta^*}^T(\mathbf{x}') \tag{D.31}$$

$$\tag{D.32}$$

This leads to the following posterior over CATE, where $\mathbf{e} = [-1,1]^\top$:

$$\tilde{\tau}(x) \sim \mathcal{N}(m^{\text{post}}(\mathbf{x})^\top \mathbf{e}, \mathbf{e}^\top \mathbf{K}_{\eta^*}^{\text{post}}(\mathbf{x},\mathbf{x}')\mathbf{e}) \tag{D.33}$$

D.2.3.1 Expected Information Gain

First, let $\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)}$ and $\mathbf{y}_e^{(1)}, \mathbf{y}_e^{(0)}$ be the covariance and outcomes for environment e that is treated and untreated respectively. To avoid confusion with treatment we will use $\mathbf{X}_{e^*}$ to refer to the host environment for this derivation. We will also use $\mathbf{K}_{|\mathscr{D}_0}$ to refer to the posterior kernel. Now to derive the Expected Information Gain in closed form we need the distribution of the following vector:

$$\begin{bmatrix} \mathbf{y}_e^{(1)} \\ \mathbf{y}_e^{(0)} \\ \tilde{\tau}(\mathbf{X}_{e^*}) \end{bmatrix} |\mathbf{X}_e, \mathscr{D}_0 \sim \mathcal{N}(\mathbf{m}, \Sigma) \tag{D.34}$$

Where we have that:

$$\Sigma_1 = \begin{bmatrix} \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)})) \\ \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(1)}) & \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{D.35}$$

$$\Sigma_2 = \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(0)}) \tag{D.36}$$

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_2 \end{bmatrix} \text{ where } \Sigma_{12} = \begin{bmatrix} \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_e^{(0)}) - \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_e^{(0)}) \\ \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_e^{(1)}) - \mathbf{K}_{|\mathscr{D}_0}(\mathbf{X}_{e^*}^{(0}, \mathbf{X}_e^{(1)}) \end{bmatrix} \tag{D.37}$$

Now from the covariance matrix we can derive the standard Expected Information Gain $\text{EIG}_{\theta|\mathscr{D}_0}$ and CATE-specific $\text{EIG}_{\theta_{c|\mathscr{D}_0}}$. Throughout we will use $\mathbf{K}$ to the posterior kernel irrespective if it has been fit or not.

Expected Information Gain over the conditional outcome parameters For the $\text{EIG}_{\mathbf{f}}$ we use the BALD form:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathscr{D}_0) - H(\mathbf{y}|\mathbf{X}_e, \mathbf{f}, \mathscr{D}_0) \tag{D.38}$$

(a) Before merging (b) After merging

Figure D1: Causal structure for the illustrative experiment.
Before merging: causal structure in D_{host}, or D_{twin} or D_{comp} *taken separately*. E acts as a collider and thus creates a dependency between X and T. After merging: causal structure in $D_{host} \cup D_{comp}$

Using the standard form of entropy for a Gaussian distribution we can read $H(\mathbf{y}_e|\mathbf{X}_e, \mathscr{D}_0)$ off of the covariance matrix above as:

$$H(\mathbf{y}_e|\mathbf{X}_e, \mathscr{D}_0) = \frac{n_e}{2}(1+\log(2\pi)) + \frac{1}{2}(\log|(\Sigma_1|)) \tag{D.39}$$

$$\text{where } \Sigma_1 = \begin{bmatrix} \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(0)}) + \sigma_0^2 I_{n_e^0} & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(0)})) \\ \mathbf{K}_\eta(\mathbf{X}_e^{(0)}, \mathbf{X}_e^{(1)}) & \mathbf{K}_\eta(\mathbf{X}_e^{(1)}, \mathbf{X}_e^{(1)}) + \sigma_1^2 I_{n_e^1} \end{bmatrix} \tag{D.40}$$

And $H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathscr{D}_0)$ being:

$$H(\mathbf{y}_e|\mathbf{f}, \mathbf{X}_e, \mathscr{D}_0) = \frac{n_e}{2}(1+\log(2\pi)) + \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{D.41}$$

This gives the Expected Information Gain as:

$$\mathscr{I}_{\mathbf{f}}(e) = \frac{1}{2}\log|(\Sigma_1|) - \frac{1}{2}\left(n_e^{(0)}\log(\sigma_0^2) + n_e^{(1)}\log(\sigma_1^2)\right) \tag{D.42}$$

Expected Information Gain on the CATE parameters We now target an Expected Information Gain on the CATE parameters on our host dataset, so $\tilde{\tau}(\mathbf{X}_0)$. By using the fact that the Expected Information Gain can be written as the mutual information between $\tilde{\tau}(X_0)$ and the observed outcomes in dataset e, we use the closed form mutual information for Gaussian's to directly write this as:

$$\mathscr{I}_{\tilde{\tau}(\mathbf{X}_0)}(e) = \frac{1}{2}\log\left(\frac{|\Sigma_1||\Sigma_2|}{|\Sigma|}\right) \tag{D.43}$$

$$\text{where } \Sigma_2 = \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(1)}) + \mathbf{K}_\eta(\mathbf{X}_{e^*}^{(0)}, \mathbf{X}_{e^*}^{(0)}) - 2\mathbf{K}_\eta(\mathbf{X}_{e^*}^{(1)}, \mathbf{X}_{e^*}^{(0)}) \tag{D.44}$$

D.3 Experimental Details

General experimental settings and hyperparameters All standard deviations and precisions were taken equal to 1. Throughout experiments, priors were taken as zero-valued vector.

In the **illustrative experiment**, 400 samples were used for computing outer expectations, and 800 samples for inner expectations. Here, we consider $X = (X_0, X_1, X_2) \in \mathbb{R}^3$. We use the sampling selection function $P_{\text{host}}(x,t) = \text{sigmoid}(1 + 2 \times x_0 - x_1 + 2 \times t) + \varepsilon$ and outcome model $y = 1 + x_0 - x_1 + x_2 + 5 \times t + 2 \times x_0 + 2 \times x_0 - 4 \times x_2 + \varepsilon$ with $X_0 \sim \mathscr{B}(12,3)$, $X_1 \sim \mathscr{N}(4,1)$, $X_2 \sim \mathscr{B}(1,7)$ and $\varepsilon \sim \mathscr{N}(0,1)$.

In both **ranking experiments**, the selection functions are randomly generated. We first generate a binary vector of the size of the dimension of X to define the subset of covariates that would be impact selection. We then generate two other random vectors, one for the multiplicative coefficients for each selected covariate, and another to define a power for each term in the sum. Ultimately, the probability of selection is taken as the sigmoid of this randomly generated polynomial.

In the **ranking experiment with IHDP**, 10 candidates were generated with a sample size ranging from 300 to 500. The host sample size was equal to 400. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 250, and a shrinkage $\alpha = 0.95$.

In the **ranking experiment with Lalonde**, 15 candidates were generated with a sample size ranging from 200 to 400. The host sample size was equal to 600. The experiment was across 20 seeds. We kept a minimum of 50 subjects in each treatment group. The hold out test dataset had a sample size of 2000. For the linear model, X, T and $X \times T$ were included. For the Gaussian Process model, a maximum of 1000 iterations was set. In CBF, both the predictive and conditional models were used with a maximum depth of 200, and a shrinkage $\alpha = 0.9$.

Datasets We describe the two datasets used in our experiments, with high-level summary given in Table D1.

	ihdp	**lalonde**
Number of samples	747	16,177
Number of features	24	8

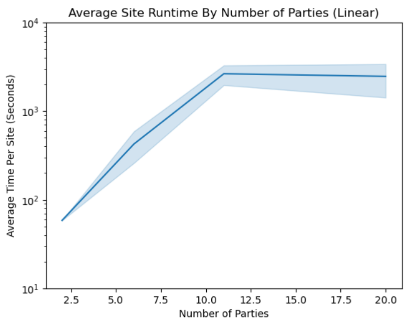
Table D1: Description of the datasets: Lalonde [185] and IHDP [169].

The **Infant Health and Development Program, or IHDP** is a randomised controlled study designed to assess how home visits by specialist doctors impact the cognitive test scores of premature infants. Initially, the dataset serves as a benchmark for evaluating treatment effect estimation algorithms, as described in [179]. This evaluation introduces selection bias by excluding non-random subsets of treated individuals to construct an observational dataset, with outcomes derived from the original covariates and treatments. The **Lalonde** originates from the National Supported Work Demonstration used by Dehejia and Wahba [254] to evaluate propensity score matching methods. The data consists of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator. The outcome is the real earnings in the year 1978.

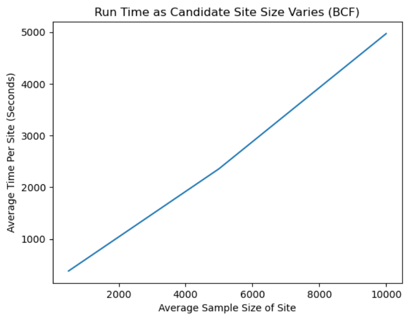
Compute times Approximate compute times for the ranking experiment on causal benchmark datasets are given in Table D2. Experiments were performed on an Apple M3 chip with a 12-core CPU and 18 GB of RAM.

	ihdp	**lalonde**
Polynomial	< 1 min	< 1 min
Causal GP	3 mins	3 mins
BART	14 mins	9 mins

Table D2: Approximate compute times.



(a) In this we repeat the multi-party computation experiments from Section 6.5.3 varying the number of sites. Each site is treated as a separate party in the multi-party computation and the average runtime per site is reported. Runtime recorded on an M3 macbook. Averaged over 5 runs.



(b) Runtime of Bayesian Causal Forest as the candidate sample size increases. We run the algorithm for the task of selecting between 5 sites with average size 500, 5000, and 10,000.

	$\rho(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial				
$\mathrm{EIG}_{\theta_c \mid \mathscr{D}_0}$	$\mathbf{0.70 \pm 0.08}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.78 \pm 0.04}$
$\mathrm{EIG}_{\theta \mid \mathscr{D}_0}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.70 \pm 0.06}$	$\mathbf{0.76 \pm 0.04}$
PropScore Error	0.40 ± 0.11	0.40 ± 0.15	0.60 ± 0.15	0.66 ± 0.04
Sample Size	0.34 ± 0.08	0.10 ± 0.08	0.27 ± 0.04	0.50 ± 0.06
CovDist	0.03 ± 0.07	0.10 ± 0.08	0.23 ± 0.06	0.48 ± 0.04
Causal GP				
$\mathrm{EIG}_{\tilde{\tau}(X_0) \mid \mathscr{D}_0}$	$\mathbf{0.49 \pm 0.06}$	$\mathbf{0.50 \pm 0.15}$	$\mathbf{0.50 \pm 0.08}$	$\mathbf{0.62 \pm 0.06}$
$\mathrm{EIG}_{\mathbf{f} \mid \mathscr{D}_0}$	0.33 ± 0.06	0.30 ± 0.15	0.43 ± 0.05	$\mathbf{0.60 \pm 0.04}$
Sample Size	0.31 ± 0.12	0.10 ± 0.20	0.20 ± 0.07	0.46 ± 0.05
PropScore Error	0.21 ± 0.09	0.10 ± 0.20	0.30 ± 0.07	0.43 ± 0.05
CovDist	0.03 ± 0.06	0.10 ± 0.20	0.160 ± 0.04	0.46 ± 0.05
Bayesian CF				
$\mathrm{EIG}_{\theta_c \mid \mathscr{D}_0}$	$\mathbf{0.54 \pm 0.10}$	$\mathbf{0.60 \pm 0.15}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
$\mathrm{EIG}_{\theta \mid \mathscr{D}_0}$	0.36 ± 0.10	0.30 ± 0.14	0.50 ± 0.07	$\mathbf{0.66 \pm 0.05}$
PropScore Error	0.45 ± 0.11	$\mathbf{0.60 \pm 0.14}$	$\mathbf{0.63 \pm 0.08}$	$\mathbf{0.70 \pm 0.04}$
Sample Size	0.16 ± 0.09	0.20 ± 0.11	0.26 ± 0.06	0.52 ± 0.05
CovDist	0.07 ± 0.09	0.00 ± 0.00	$0.26 \pm 0.06a$	0.46 ± 0.05

Table D3: IHDP dataset ranking experiment results with 10 candidate datasets

D.4 Further experimental results

For completeness, we include the performance of all baselines on the IHDP dataset in Table D3 and the Lalonde in Table D4.

	$\rho(\uparrow)$	**p@1** $(\uparrow)$	**p@3** $(\uparrow)$	**p@5** $(\uparrow)$
Polynomial				
$\text{EIG}_{\theta_c \mid \mathscr{D}_0}$	$\mathbf{0.47 \pm 0.05}$	$\mathbf{0.40 \pm 0.11}$	$\mathbf{0.60 \pm 0.05}$	$\mathbf{0.79 \pm 0.04}$
$\text{EIG}_{\theta \mid \mathscr{D}_0}$	$\mathbf{0.43 \pm 0.05}$	$\mathbf{0.35 \pm 0.13}$	0.48 ± 0.04	0.53 ± 0.03
PropScore Error	0.19 ± 0.10	0.20 ± 0.17	0.32 ± 0.06	0.49 ± 0.05
Sample Size	0.24 ± 0.10	0.25 ± 0.08	0.38 ± 0.08	0.58 ± 0.06
CovDist	0.20 ± 0.04	0.25 ± 0.07	0.38 ± 0.06	0.48 ± 0.07
Causal GP				
$\text{EIG}_{\tilde{\tau}(X_0) \mid \mathscr{D}_0}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.5 \pm 0.12}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.72 \pm 0.03}$
$\text{EIG}_{\mathbf{f} \mid \mathscr{D}_0}$	$\mathbf{0.41 \pm 0.04}$	$\mathbf{0.4 \pm 0.1}$	0.43 ± 0.04	0.58 ± 0.07
PropScore Error	0.19 ± 0.05	0.21 ± 0.15	0.32 ± 0.06	0.43 ± 0.07
Sample Size	0.13 ± 0.07	0.15 ± 0.08	0.31 ± 0.09	0.53 ± 0.06
CovDist	0.22 ± 0.04	0.25 ± 0.07	0.36 ± 0.08	0.48 ± 0.04
Bayesian CF				
$\text{EIG}_{\theta_c \mid \mathscr{D}_0}$	$\mathbf{0.44 \pm 0.05}$	$\mathbf{0.45 \pm 0.08}$	$\mathbf{0.55 \pm 0.05}$	$\mathbf{0.78 \pm 0.04}$
$\text{EIG}_{\theta \mid \mathscr{D}_0}$	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.45 \pm 0.06}$	0.43 ± 0.04	0.52 ± 0.03
PropScore Error	0.22 ± 0.06	0.2 ± 0.07	0.32 ± 0.06	0.47 ± 0.07
Sample Size	0.18 ± 0.07	0.2 ± 0.06	0.35 ± 0.09	0.41 ± 0.06
CovDist	0.34 ± 0.03	0.3 ± 0.07	0.42 ± 0.08	0.61 ± 0.04

Table D4: Lalonde dataset ranking experiment results with 15 candidate datasets

D.5 Related work: further details

On Causal Federated Learning Federated learning is a distributed machine learning approach that enables multiple parties to collaboratively train a shared model while keeping their raw data decentralised and private. Various federated learning approaches have been proposed, including federated stochastic gradient descent [255], federated averaging [256], and more recently, methods for joint learning of deep neural network models [257, 258]. However, these algorithms do not inherently support causal inference, as the dissimilar distributions across different data sources may lead to biased causal effect estimation. To date, limited research has been conducted on the federated estimation of causal effects, highlighting the need for further exploration in this area. Due to the scope of our work, in the following paragraphs, we will focus on presenting Federated Learning methods for CATE estimation, where covariate distribution and treatment allocation are not assumed to be identical across datasets.

Several methods propose disentangling the loss function to facilitate federated learning. Vo et al. [194] propose CausalRFF, an adaptive kernel approach for causal inference that utilises Random Fourier Features to partition the loss function into multiple components, with each component corresponding to a specific data source. However, CausalRFF approach lacks strong privacy guarantees to prevent data recovery, and modelling complex non-linear relationships remains challenging [197]. Liu et al. [195] introduce a Bayesian method where parameters refer to a local disentangled loss and are updated cross-silo using server aggregation. Similarly, Vo et al. [196] divide the loss function into site-specific functions, and specify a variational posterior distribution for each local loss.

Instead of tackling the loss function, Almodóvar et al. [197]) introduce a method based on disentanglement of latent factors into instrumental, confounding, and risk factors, which are then used for treatment effect estimation. They apply federated averaging on a neural network-based generative causal inference model. Ultimately, FedCov [259] is a parametric method for federated adjustment of covariate distributions between sites, where sample weights are derived from a propensity-like model.

In all the aforementioned methods, the accuracy of causal estimation is reduced due to the constraints of federated learning. Conversely, our approach does not alter the causal estimation step, thereby maintaining optimal estimation accuracy. The framework we propose focuses on federated learning of the Expected Information Gain that would be obtained by merging with a dataset. While some Federated Causal Learning methods [194, 196] provide uncertainty bounds, which could potentially be used to decide which dataset to merge with by comparing the uncertainty in these bounds, the provided bounds apply to the federated estimate and not the causal estimate potentially obtained after merging. Ultimately, none of these methods provide strong privacy guarantees, such as differential privacy (DP), which would ensure that raw data cannot be recovered from the model parametrisation or summary statistics. Moreover, all these methods use the outcome values for training their federated model, and outcome values tend to be more sensitive in nature.

On Causal Differential Privacy Contrasting with previous approaches, Niu et al. [198] introduce a meta-algorithm that adds differential privacy (DP) guarantees to various popular CATE estimation frameworks, addressing the privacy concerns mentioned earlier. However, their method relies on multiple sample splitting, where separate subsets of the data are used for estimating the propensity score and the joint response model. This approach allows for parallel composition, a property of differential privacy. In contrast, our work prioritises data efficiency, and aims to utilise the entire dataset for CATE estimation without the need for sample splitting.

On Bayesian Experimental Design Bayesian Active Learning by Disagreement (BALD) [172] is a method designed to strategically acquire training data by focusing on regions of high uncertainty. BALD introduces an acquisition function rooted in information theory, which guides the data acquisition process. When reducing entropy towards all parameters in 6.3.1, we introduce a new setting for BALD where dataset are considered as data points. In the CausalBALD [199] approach, the acquisition function is altered to specifically target areas where the distributions of different treatment groups overlap, thereby maximizing sample efficiency for learning personalised treatment effects. CausalBALD is also made for the acquisition of individual data points. However, contrarily to BALD, CausalBALD's acquisition function cannot provide a scalar measure if we compute it for a dataset (i.e. a matrix $\{\mathbf{x}_i, t_i\}_{i=1}^{n_e}$) instead of data points (i.e. a vector $\mathbf{x}_i, t_i$ for a given i). To apply CausalBALD in our setting, one would need to approximate the higher-order interaction terms between all combinations of data points within each dataset, thus making the computation intractable.