cc-Shapley: Measuring Multivariate Feature Importance Needs Causal Context

Jörg Martin[1] **Stefan Haufe**[1,2,3]

[1]Physikalisch-Technische Bundesanstalt, Berlin, Germany
[2]Technische Universität, Berlin, Germany
[3]Charité Universitätsmedizin, Berlin, Germany

Abstract

Explainable artificial intelligence promises to yield insights into relevant features, thereby enabling humans to examine and scrutinize machine learning models or even facilitating scientific discovery. Considering the widespread technique of Shapley values, we find that purely data-driven operationalization of multivariate feature importance is unsuitable for such purposes. Even for simple problems with two features, spurious associations due to collider bias and suppression arise from considering one feature only in the observational context of the other, which can lead to misinterpretations. Causal knowledge about the data-generating process is required to identify and correct such misleading feature attributions. We propose cc-Shapley (causal context Shapley), an interventional modification of conventional observational Shapley values leveraging knowledge of the data's causal structure, thereby analyzing the relevance of a feature in the causal context of the remaining features. We show theoretically that this eradicates spurious association induced by collider bias. We compare the behavior of Shapley and cc-Shapley values on various, synthetic, and real-world datasets. We observe nullification or reversal of associations when moving from observational to cc-Shapley.

1 INTRODUCTION

Feature attribution is a common paradigm in the field of Explainable Artificial Intelligence (XAI). Its goal is to quantify the importance of each feature by attributing a score to it. It is often assumed that using such techniques allows us to check whether a certain model is "correct", i.e., whether it misbehaves on certain data points or whether it uses unexpected, possibly unfavorable, structures for its prediction [Ribeiro et al., 2016, Lapuschkin et al., 2019, Anders et al., 2022]. Likewise, XAI is hoped to boost scientific discovery by unveiling hidden multivariate patterns in the data that are associated with prediction targets of interest [Jiménez-Luna et al., 2020, Wong et al., 2024, Samek and Müller, 2019]. All of these promises are based on the assumption that the attributions gained with such methods do not show spurious patterns and can be interpreted unambiguously.

This work aims to show that this assumption is an illusion unless we include causal knowledge into our considerations. We observe that a purely observational approach to XAI produces spurious relevance patterns that can easily lead to false conclusions. As these patterns are rooted in the causal relationship of the involved variables, causal knowledge is required to correct them.

Our analysis focuses on Shapley values, a concept of game-theoretical origin that has become a well-established approach to XAI. When predicting a target variable Y from features $\mathcal{F} = \{X_1, \dots, X_n\}$ the *Shapley value* of a feature X_j is defined as[1] [Shapley et al., 1953]

$$\phi(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!} I_{\mathcal{S}}(X_j)\,, \quad (1)$$

where we introduce the notation

$$I_{\mathcal{S}}(X_j) = \mathbb{E}[Y|X_j, \mathcal{S}] - \mathbb{E}[Y|\mathcal{S}] \quad (2)$$

for the change in prediction when we add observation X_j to the observed features $\mathcal{S} \subseteq \mathcal{F}$. Formula (1) computes the weighted sum of these changes, where the combinatorial weights can be interpreted through the probability of randomly drawing $\mathcal{S}$, cf. Appendix A. For binary $Y \in \{0, 1\}$, as considered in some examples below, (2) simplifies to a difference of probabilities.

The computational complexity behind (1) quickly increases with $|\mathcal{F}|$. For this reason, several scalable modifications

[1]There are various versions of (1), cf. Appendix A.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4323–4352.

have been proposed, e.g., by Lundberg and Lee [2017] and Janzing et al. [2020]. We show, however, that even without any approximations put on top, ϕ is unsuitable for the purposes generally targeted by XAI. To illustrate this, we will use the following easy running example throughout this work.

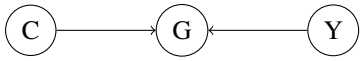
Example 1.1 (Breakfast and diabetes)**.** *Suppose blood glucose G is measured in a patient to assess whether he or she has diabetes ($Y = 1$) or not ($Y = 0$). Unfortunately, the patient ignores the doctor's request not to eat breakfast before the test. We assume that the measured value G follows the (simplified) relation*

$$G = 85 + 0.4 \cdot C + 40 \cdot Y + U\,, \tag{3}$$

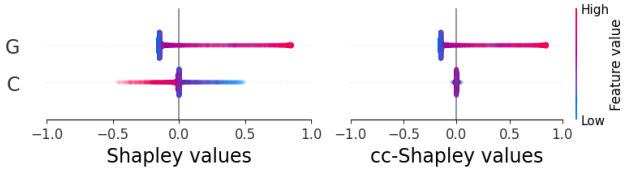
where C denotes the carbohydrate intake during breakfast and U denotes independent noise. We assume that $Y \sim \text{Bernoulli}(0.15)$, $C \sim \mathcal{N}(60; 25^2)$ and $U \sim \mathcal{N}(0; 10^2)$.

The graph in Figure 1a illustrates the causal structure of Example 1.1: the value of G is caused by the variables Y and C. The plot on the left of Figure 1b shows the Shapley values (1) for data generated from this simple problem. A red color indicates a high value of a feature, whereas a blue color indicates a low value. A positive Shapley value $\phi(G) = \frac{1}{2}I_\emptyset(G) + \frac{1}{2}I_C(G)$ indicates a positive association whereas larger G tend to co-occur with larger Y if G is either considered alone ($\mathcal{S} = \emptyset$) or in the context of a fixed C ($\mathcal{S} = \{C\}$). As red points in Figure 1b are concentrated on the right for G, we could conclude that high G values makes diabetes more likely, which matches intuition. We call this *positive relevance* in the following.

The Shapley values for C, on the other hand, indicate a *negative relevance*: high values of C are associated with a lower diabetes probability. Does this imply that a single high carbohydrate intake makes it less likely that the patient has diabetes? This seems absurd and rightly so. The underlying reason is instead something which one might dub the "explain away effect" but which is typically known as *suppression* in the literature [Conger, 1974]: When the patient consumed many carbohydrates there is no need for diabetes to explain high values of G. The presence of suppression in our example can be read off the causal graph in Figure 1a: the suppressor C is connected to the target Y via a node G with two incoming arrows (known as a *collider*). Conditioning on a collider leads to a spurious association, known as *collider bias*, which makes C a suppressor of G. Collider bias and suppression are in this work considered as two sides of the same coin, cf. Section 2.2. Not every feature attributed negative relevance by ϕ will automatically be a suppressor. In other examples a feature can, of course, be of negative relevance as it really causes the target to decrease. Apparently, the "right" interpretation of a chart like the left one in Figure 1b can vary from situation to situation. This ambiguity can easily lead to wrong conclusions and



(a) Causal graph for Example 1.1



(b) Shapley (left) and cc-Shapley values (right) for Ex. 1.1

Figure 1: Causal graph for Example 1.1 (top), together with the according results for conventional Shapley values (1) (bottom - left) and the cc-Shapley values (6) (bottom - right). Experimental details are in Appendix E.1.

undermines the usability of XAI for purposes such as model analysis, let alone for scientific discovery.

Why were we able to tell that a suppression effect rather than a potential health benefit of one carbohydrate-rich breakfast is a more likely interpretation for the negative relevance of C in Figure 1b? We had to use something that a purely data-driven model does not have: an understanding of the real world, encoded in the causal diagram in Figure 1a. In this article we propose to use this understanding for our treatment of the context variables $\mathcal{S}$ in (1) via techniques from *causal inference* [Pearl, 2009]. This causal modification of (1) is called *cc-Shapley* values, and is introduced in Definition 3.1. For Example 1.1, we depict the cc-Shapley values on the right side of Figure 1b. We observe that the cc-Shapley values do not attribute C any importance for diabetes risk, which matches the intuition.

The main contributions of this article are as follows:

- We highlight a fundamental problem underlying non-causal XAI methods such as (1). In causal inference, this problem is known as collider bias or suppression.
- We propose cc-Shapley values, which modify (1) to use causal information. This is, to the best of our knowledge, the first approach designed to avoid collider bias without the need to restrict oneself to univariate feature importance.
- We present theoretical and experimental results regarding the computability and behavior of cc-Shapley values in the presence of collider bias. We use several synthetic and one real world example.

The article is structured as follows: In Section 2 we recall some background on causal inference including collider bias and suppression, which are the central topics of this work. In Section 3 we introduce cc-Shapley values and give an algorithm for their estimation. Section 3.3 lists the main limitations of this work, such as the restriction to tabular data and the requirement to have a causal graph. In Section

4 we compare the results of Shapley and cc-Shapley values on various examples.

Comparison With the Literature. There are a plethora of XAI methods in the literature, cf. for example the reviews by Angelov et al. [2021], Minh et al. [2022] and Mersha et al. [2024]. While this work focuses entirely on Shapley values, the criticism raised here goes beyond this specific method. Typically, non-causal methods use only observational ingredients: the data and/or the model (which is also trained from the data). As causal information is not recoverable from observations alone in general and as collider bias is a causal phenomenon, it is implausible that any combination of purely observational information can eradicate the influence of collider bias. Indeed, Wilming et al. [2023] demonstrate the susceptibility of many popular XAI methods to suppression.

There is a growing awareness that the current approach to XAI needs scrutiny or even revision [Freiesleben and König, 2023, Haufe et al., 2024, Wilming et al., 2022] with some works discussing the need and possibility for combining XAI with causal concepts, e.g., [Carloni et al., 2025, Beckers, 2022, Karimi, 2025, Holzinger et al., 2019, Schölkopf, 2022]. Various works propose using the underlying causal structure to produce counterfactual explanations, see, e.g., the works by Karimi et al. [2021, 2020], König et al. [2023]. The focus in these works is mostly to produce counterfactuals that are coherent with the actual world, but the topic of suppression is not discussed in this context.

Janzing et al. [2020], Budhathoki et al. [2022], Heskes et al. [2020], Frye et al. [2020] and Jung et al. [2022] also discuss variants of Shapley values invoking causal concepts. We discuss the relation of our method to these variants in Section 3.1 and, in greater detail, in Appendix B. Wilming et al. [2023] study (conventional) Shapley values for a simple suppressor problem for which no relevance is attributed to a suppressor; this statement is however only true for their considered example and a specific approximation to Eq. (1). Haufe et al. [2014] and Clark et al. [2025] discuss the problem of suppression and observe that a univariate approach to feature importance avoids these problems. We will discuss the relation of our work to this idea in Section 2.2 including a discussion why univariate importance is in general not sufficient to assess the relevance of a feature.

2 BACKGROUND

2.1 BASICS OF CAUSAL INFERENCE

We here recall core concepts of causal inference needed in this article and refer to Pearl [2009]. In causal inference, the data genererating process behind a set of variables is described by a *structural causal model (SCM)*, which is a triplet $\mathcal{M} = (\mathcal{V}, (f_X)_{X\in\mathcal{V}}, (U_X)_{X\in\mathcal{V}})$, where

- $\mathcal{V}$ is the set of variables,
- $(U_X)_{X\in\mathcal{V}}$ is a set of independent[2] random variables, representing noise, and
- each variable $X \in \mathcal{V}$ is related to its *parents* $\mathrm{pa}_X \subseteq \mathcal{V}\setminus\{X\}$ and its noise variable U_X via an *assignment function* $X = f_X(\mathrm{pa}_X, U_X)$.

To avoid inconsistencies, the relations are assumed to be acyclic: That is, when drawing arrows from pa_X to X for each $X \in \mathcal{V}$, we arrive at a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edges $\mathcal{E}$. Given an instance of the noise variables $(U_X)_{X\in\mathcal{V}}$, we can find the values of all $X \in \mathcal{V}$ through evaluation of the assignment functions, starting from the source nodes without parents. The graph $\mathcal{G}$ is called the *causal graph* and represents the causal functional dependencies underlying the data generating process. Figure 1a depicts the causal graph for Example 1.1.

Statistical association between variables can be assessed by studying the paths in $\mathcal{G}$. A *path* between two variables is a concatenation of adjacent edges from $\mathcal{E}$ (independent of the direction of the edges) leading from one variable to the other. If all edges in a path have the same direction, we denote it by $\dashrightarrow$ or $\dashleftarrow$ and call it a *causal path*. If a causal path consists of only one edge we call it *direct*, otherwise *indirect*. A path between two sets $\mathcal{S}$ and $\mathcal{S}'$ is called a *backdoor path* if it leaves $\mathcal{S}$ with an incoming edge. A path is called *blocked*, conditional on a (possibly empty) set $\mathcal{Z} \subseteq \mathcal{V}$, if it

- contains a $Z \in \mathcal{Z}$ that acts on this path as a *chain* $\to Z \to$ (or $\leftarrow Z \leftarrow$) or as a *fork* $\leftarrow Z \rightarrow$, or
- if it contains a *collider* $\to C \leftarrow$ which is neither contained in $\mathcal{Z}$ nor in the *ancestors* $\mathrm{an}(\mathcal{Z})$ of $\mathcal{Z}$ (the set of nodes that have a causal path leading to $\mathcal{Z}$).

A path is called *unblocked* if it is not blocked. If all paths between two sets of variables $\mathcal{S}, \mathcal{S}' \subseteq \mathcal{V}$ are blocked conditional on $\mathcal{Z}$, we call them *d-separated* and write $\mathcal{S} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{S}'|\mathcal{Z}$. d-separation implies independence: $\mathcal{S} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{S}'|\mathcal{Z} \Rightarrow \mathcal{S} \perp\!\!\!\perp \mathcal{S}'|\mathcal{Z}$. The graph $\mathcal{G}$ is called *faithful* if the reverse direction is always true.

Intervention with the do-operator is the process of creating from $\mathcal{M} = (\mathcal{V}, (f_X)_{X\in\mathcal{V}}, (U_X)_{X\in\mathcal{V}})$ a new SCM $\mathcal{M}^{\mathrm{do}(X'=x')} = (\mathcal{V}, (f_X)_{X\in\mathcal{V}\setminus\{X'\}} \cup \{\tilde{f}_{X'}\}, (U_X)_{X\in\mathcal{V}})$ where the assignment function $f_{X'}(\mathrm{pa}_{X'}, U_{X'})$ is replaced with the constant function $\tilde{f}_{X'}(U_{X'}) = x'$ (with a dummy $U_{X'}$-dependency). The distribution in this modified model is denoted by $P(\,\cdot\,|\mathrm{do}(X' = x'))$ or simply $P(\,\cdot\,|\mathrm{do}(X'))$ to indicate that X' is set to a some value x' which we do not need to specify further. Intervention simulates the effect of a variable when we cut off its usual causes and corresponds to deleting all incoming arrows to X' in $\mathcal{G}$. Similarly,

[2]Throughout this work, we assume that the SCM satisfies the local Markov property, i.e., there are no hidden confounders outside the set $\mathcal{V}$.

we can define an intervention $\mathrm{do}(\mathcal{X})$ on a subset of variables $\mathcal{X} \subseteq \mathcal{V}$ or a stochastic intervention $\mathrm{do}(\mathcal{X} \sim q)$ when drawing the values $\mathcal{X}$ from a distribution q instead of using constant values [3].

In supervised machine learning, the typical task is to estimate a target variable from other variables. To account for this distinction in our setup, we call one variable $Y \in \mathcal{V}$ the *target* and the remaining variables $\mathcal{F} = \mathcal{V} \setminus \{Y\}$ *features*. The features will be indexed as $\mathcal{F} = \{X_i : 1 \leq i \leq n\}$.

2.2 COLLIDER BIAS

Suppressor Variables and Collider Bias. In Figure 1b, we observed that even for simple problems, uninformative features, i.e., features that are independent of the target, can be attributed importance by XAI methods. This is because these features might be useful to remove variance from informative features. In the literature, variables that are useful in such a secondary sense are known as *suppressor variables*, as they can be used to suppress noise [Horst et al., 1941, Conger, 1974, Darlington, 1968, Kim, 2019]. These references all give different definitions of what a suppressor is, partially contradicting each other or giving definitions that only make sense in a linear framework. We here follow Wilming et al. [2022] and Haufe et al. [2024] and define suppression through *collider bias*, the statistical association that arises due to the unblocking of paths when conditioning on a collider or its ancestors (cf. Section 2.1). We call a (possibly informative) feature a *suppressor* if it is connected to the target via paths that contain colliders. In this view, suppression and collider bias are just two sides of the same coin: suppressors are those variables whose association with the target might be affected by collider bias when conditioning on other variables. Example 1.1 with the causal graph from Figure 1a is a very easy example on how the collider G makes C a suppressor. For general setups and a singleton $\mathcal{S} = \{X_k\}$ the middle column of Table 1 summarizes how conditioning on X_k can lead to collider bias between a feature X_j and the target, making X_j a suppressor. The rows of Table 1 distinguish various types of paths τ between X_j and Y. In cells marked in red, conditioning on X_k will (potentially) lead to collider bias between X_j and Y: τ was previously blocked but might become unblocked inducing or distorting statistical association. As the number of paths quickly grows with $|\mathcal{F}|$, the consequences of spurious association introduced by such effects is likely to become aggravated even for low-dimensional problems.

Univariate Importance Is Insufficient. The problem of suppressor variables and the susceptibility of many XAI methods to them is known in the XAI literature [Wilming et al., 2022, Haufe et al., 2024, König et al., 2025, Weichwald et al., 2015]. Haufe et al. [2014] observed that even the weights of linear models will highlight suppression variables as relevant. Clark et al. [2025] and Gjølbye et al. [2025] extended this observation to generalized additive models and locally linear explanations. These three works also propose potential remedies for this problem, which could, in a nutshell, be summarized as (re-)fitting the considered feature X_j to the target Y in a univariate fashion[4]. In the notation of (2), we can express such an approach as

$$I_\emptyset(X_j) = \mathbb{E}[Y|X_j] - \mathbb{E}[Y]\,, \tag{4}$$

where we recall that the conditional expectation is the optimal fit of X_j onto Y, cf. Appendix C.1. The object (4) is not susceptible to collider bias and thus to suppression as it does not condition on other features. In particular, we have $X_j \perp\!\!\!\perp Y \Rightarrow I_\emptyset(X_j) = 0$ so that (4) only highlights features that are statistically associated with Y, a property called the *statistical association property* (SAP) by Clark et al. [2025], Wilming et al. [2023] and Haufe et al. [2024]. For Example 1.1, the SAP guarantees that $I_\emptyset(C) = 0$.

Univariate importance measures such as (4) avoid attributing importance to statistically irrelevant features which is necessary to apply XAI for model debugging and scientific discovery. However, the interplay between variables is vital to the performance of a ML model and interactions of variables can create information that is not available by looking at each variable on its own as shown by

Example 2.1 (Univariate importance is not enough)**.** *Let* $X_1, X_2 \sim \mathrm{Bernoulli}(\frac{1}{2})$ *be independent and* $Y := X_1 \oplus X_2$ *with* $\oplus$ *denoting the XOR operation. Then* $I_\emptyset(X_1) = I_\emptyset(X_2) = 0$ *indicate no importance of the features whereas the "higher order" importance terms* $I_{X_2}(X_1) = X_1 \oplus X_2 - \frac{1}{2}$, $I_{X_1}(X_2) = X_1 \oplus X_2 - \frac{1}{2}$ *(correctly) indicate importance.*

A more complete XAI method will thus have to consider the multivariate interplay between variables. As we saw above, doing this purely based on observational data as done by mainstream approaches can lead to spurious association that can easily be misinterpreted. In the following section we demonstrate that using causal information provides a way out of this dilemma.

[3] To define stochastic interventions we actually have to allow in general for dependencies between exogenous variables; for all cases studied here, this technicality is, however, irrelevant. We discuss this and stochastic interventions in more detail in Section C.4 in the appendix.

[4] Haufe et al. [2014] actually propose a method that, under the assumption of Gaussianity, is equivalent to a univariate linear regression from target to feature. The regression coefficient of such a fit equals, up to standard deviations, a linear univariate regression from feature to target.

Paths τ between X_j and Y	$\dots \vert X_k$	$\mathrm{do}(X_k)$
$\to X_k \to$ exists on τ	✗ or ✓⇒ ✗	✗ or ✓⇒ ✗
$\leftarrow X_k \to$ exists on τ	✗ or ✓⇒ ✗	✗ or ✓⇒ ✗
$\to X_k \leftarrow$ exists on τ	✗⇒ ✗ or ✓	✗ ⇒ ✗
$\to C \leftarrow$ exists on τ and $X_k \notin \tau, C \in \mathrm{an}(X_k)$	✗⇒ ✗ or ✓	✗ ⇒ ✗
None of the above	no effect	no effect

Table 1: Change in the state of being blocked (✗) or unblocked (✓) for paths τ between feature X_j and target Y when conditioning (middle column) or intervening (right column) on X_k. "✗ or ✓" means that the path is potentially unblocked - depending on other path details. Conditioning on X_k potentially unblocks previously blocked paths, which can lead to *collider bias* (cells marked in red). This is avoided by causal interventions $\mathrm{do}(X_k)$.

3 METHODOLOGY

3.1 USING THE CAUSAL CONTEXT

A Causal Version of Shapley Values. We saw in Section 1 for our running Example 1.1 that, even in this simple case, objects such as $\mathbb{E}[Y|C, G]$ are susceptible to the problem of spurious correlations, namely between C and Y conditional on G. We here propose to solve problems of this kind as described in the following.

First, abandoning the symmetry between the considered features, we propose to distinguish in each scenario between

1. The variable whose importance is to be studied.
2. The variable(s) that describe the background within which the importance of the variable from 1 is studied. We call these variables the *context*.

The "paradoxical" behavior for Example 1.1 arose when studying the importance of the feature C (carbohydrate intake) in the context of a glucose value of G. The quantity $\mathbb{E}[Y|C, G]$ considers the collider G as an observation, which leads to the spurious correlation between the suppressor C and Y. We here propose to apply *interventions* on the context variable(s) instead of conditioning on observed values.

Definition 3.1. *We define the* importance of a feature X_j in the causal context of features $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ *as*

$$I_{\mathrm{do}(\mathcal{S})}(X_j) = \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(\mathcal{S})]\,. \quad (5)$$

Using (5) *we define causal context Shapley (*cc-Shapley*) values as*

$$\phi_{cc}(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!} I_{\mathrm{do}(\mathcal{S})}(X_j)\,. \quad (6)$$

Intervening on the context allows us to remove collider bias: For the case of a univariate context $\mathcal{S} = \{X_k\}$, the rightmost column of Table 1 summarizes the effect of the intervention $\mathrm{do}(X_k)$ on the blocking of paths between X_j and Y. In contrast to a mere conditioning on X_k (middle column), the intervention $\mathrm{do}(X_k)$ does not unblock a path τ that was previously blocked. In fact, the same holds for general context variables $\mathcal{S}$: Since intervention removes incoming arrows to $\mathcal{S}$ in $\mathcal{G}$, the operation $\mathrm{do}(\mathcal{S})$ does not lead to a conditioning on colliders.

For the running Example 1.1, we obtain, from Lemmas 3.3, 3.4 below and Appendix E.1, $I_{\mathrm{do}(C)}(G) = I_C(G) \simeq \sigma(\frac{2}{5}(G - 0.4C - 109)) - 0.15$ and $I_{\mathrm{do}(G)}(C) = I_\emptyset(C) = 0$. Note that the information on the suppressor C is still used but now counts towards the importance of the informative variable G it suppresses. The uninformative suppressor feature C is assigned no importance. This can be shown to hold in general for any non-informative feature whenever the underlying causal graph is faithful. Thus, the cc-Shapley values as defined in Definition 3.1 fulfill the SAP.

Proposition 3.2 (SAP for Definition 3.1)**.** *Given a feature X_j with $X_j \perp\!\!\!\perp_{\mathcal{G}} Y$ we have $I_{\mathrm{do}(\mathcal{S})}(X_j) = 0$ for any $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$. This implies that we have $X_j \perp\!\!\!\perp_{\mathcal{G}} Y \Rightarrow \phi_{cc}(X_j) = 0$ for ϕ_{cc} as in* (6)*.*

Proof. This follows from the fact that an intervention on $\mathcal{S}$ cannot unblock paths between X_j and Y, hence $X_j \perp\!\!\!\perp_{\mathcal{G}} Y|\,\mathrm{do}(\mathcal{S})$ and thus $X_j \perp\!\!\!\perp Y|\,\mathrm{do}(\mathcal{S})$. □

Comparison With Other Methods. Jung et al. [2022] introduce do-*Shapley* via the formula

$$\phi_{\mathrm{do}}(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \gamma(\mathcal{S})(\mathbb{E}[Y|\mathrm{do}(X_j, \mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(X_j)]\,.$$

This expression appears similar to the one proposed for ϕ_{cc} above, with the exception that we do not intervene on X_j in Definition 3.1. This difference is however crucial for the aspects studied in this work. Broadly speaking, prediction problems can often be classified as causal or anti-causal, depending on whether the target is the effect or the cause of the features [Schölkopf et al., 2012]. While the issue of suppression is not exclusive to the anti-causal setup, it most naturally occurs in such settings due to its connection to collider bias. In fact, almost all causal settings studied in this work (except for Example 2.1) contain at least some anti-causal elements. The object ϕ_{do} is not designed for anti-causal settings and doesn't attribute relevance to any feature when used in this context (cf. Appendix B). The same is true for the *intrinsic causal contribution* approach proposed by Janzing et al. [2024]. Heskes et al. [2020] propose *causal Shapley values* that differ from do-Shapley by using the actual model instead of Y. We find in Appendix B that this leads for Example 1.1 once more to spurious negative

relevance attributed to C. Frye et al. [2020] also use the model for their concept of *asymmetric Shapley values*. We find in Appendix B that their approach can, however, be modified to a model-agnostic form and that this form is, similar to cc-Shapley, well-behaved on Example 1.1. In the presence of mediators this modification shows however inconsistencies in its behavior on causal and anti-causal problems in contrast to cc-Shapley. For a more detailed discussion we refer to Section B of the appendix.

3.2 ESTIMATION

Simple Cases. Objects such as $\mathbb{E}[Y|\mathcal{S}]$ and $\mathbb{E}[Y|X_j, \mathcal{S}]$ can rather easily be learned from data via training a machine learning model of choice that predicts Y from $\mathcal{S}$ or $\{X_j\} \cup \mathcal{S}$, cf. Lemma C.1 in the appendix. Interventional objects such as $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$ from Definition 3.1 are harder to obtain and require identification (e.g., via do-calculus) or access to (an estimate of) the SCM. But in some cases these objects match simple observational quantities as shown by the following two lemmas whose proofs are in Appendix C.2.

Lemma 3.3 (Irrelevant context)**.** *Consider $X_j \in \mathcal{F}, \mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ and assume that there are no causal paths $\mathcal{S} \dashrightarrow Y, X_j$. Then we have $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|X_j]$ and $I_{\mathrm{do}(\mathcal{S})}(X_j) = I_\emptyset(X_j)$.*

For the running Example 1.1, Lemma 3.3 implies that $I_{\mathrm{do}(G)}(C) = I_\emptyset(C) = 0$ as there are no causal paths $G \dashrightarrow Y$ or $G \dashrightarrow C$ and hence G is irrelevant as causal context for C.

Lemma 3.4 (Intervention equals observation)**.** *Consider $X_j \in \mathcal{F}$, $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ such that, either*

- (no backdoor paths) *there are no unblocked backdoor paths from $\mathcal{S}$ to X_j or from $\mathcal{S}$ to Y, or*
- (purely causal setup) *there are no causal paths $Y \dashrightarrow X_j, \mathcal{S}$ and no confounders $H \in \mathcal{V} \setminus (\{Y, X_j\} \cup \mathcal{S})$ with $X_j \dashleftarrow H \dashrightarrow Y$ or $\mathcal{S} \dashleftarrow H \dashrightarrow Y$,*

then we have the identities $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|X_j, \mathcal{S}]$ and $I_{\mathrm{do}(\mathcal{S})}(X_j) = I_\mathcal{S}(X_j)$.

The feature C in Example 1.1 has no backdoor paths so that Lemma 3.4 implies $I_{\mathrm{do}(C)}(G) = I_C(G)$. The same argument implies that for Example 2.1 we have $I_{\mathrm{do}(X_2)}(X_1) = I_{X_2}(X_1)$ and $I_{\mathrm{do}(X_1)}(X_2) = I_{X_1}(X_2)$ and thus $\phi_{cc}(X_1) = \phi(X_2) \neq 0, \phi_{cc}(X_2) = \phi(X_2) \neq 0$.

Backdoor Adjustment. In Lemma C.5 in the appendix we provide a formula for the identification of $I_{\mathrm{do}(\mathcal{S})}(X_j)$ from observational data similar to the backdoor adjustment from Pearl [2009].

Algorithm 1: Compute $I_{\mathrm{do}(\mathcal{S})}(X_j)$ from SCM

Input: SCM $\mathcal{M}$, $X_j \in \mathcal{F}$, context $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$
Output: $I_{\mathrm{do}(\mathcal{S})}(X_j)$

```
# Create modified model
```
Create sampler of marginal $q(\mathcal{S})$ in $\mathcal{M}$;
Construct $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$;
```
# Fit ML models (NN, XGBoost,...)
```
Sample $(X_j, \mathcal{S}, Y)$ from $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$;
$\mathbb{E}[Y|\mathrm{do}(\mathcal{S})] \leftarrow$ fit $\mathcal{S}$ to Y;
$\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] \leftarrow$ fit $X_j, \mathcal{S}$ to Y;
```
# Compute importance
```
$I_{\mathrm{do}(\mathcal{S})}(X_j) \leftarrow \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$;
return $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$, $I_{\mathrm{do}(\mathcal{S})}(X_j)$

Using the SCM. Algorithm 1 summarizes how $I_{\mathrm{do}(\mathcal{S})}(X_j)$ can be computed given an SCM $\mathcal{M}$. This is the method that we use in Section 4 below. We first isolate the joint marginal q of the context variables $\mathcal{S}$ within the original model $\mathcal{M}$. We then perform a stochastic intervention on the SCM to obtain $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$. In this modified model, we can obtain estimates of the functions $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$ and $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$ via standard multivariate fitting using data-driven models, cf. Appendix C.4.

Unless one is working with synthetic examples, the SCM will naturally not be known, as we also discuss in Section 3.3 below. However, given a causal graph and a defined noise structure, the functional relations can be estimated from observational data by regressing each feature on its parents [Peters et al., 2017]. For instance, for the real world example in Section 4 we learn each assignment function via a non-linear logistic regression with the parents as input variables.

3.3 LIMITATIONS

Neither the original Shapley values (1) nor our causal analogue (6) are scalable in the way they are formulated here, cf. Section D in the appendix. The number of of context sets $\mathcal{S}$ increases exponentially with $|\mathcal{F}|$ and for each feature X_j and context $\mathcal{S}$ two data-driven models have to be fitted to estimate $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$ and $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$. Scalable approximations [Chen et al., 2023] could be a way out but were not studied in the scope of this work. [Parafita et al., 2025] introduce an estimand-agnostic approach, that builds on the insight that do-Shapley values (cf. Section B.1) are directly estimable from the joint distribution of all variables. The same statement is true for cc-Shapley, but the application of this idea to our approach was not studied here. An omission of redundant terms, as proposed by Parafita et al. [2025], Witter et al. [2026] for do-Shapley, would also enhance the computability of cc-Shapley, but appears to be more com-

plex due to the presence of conditioning along interventions in (5). We did not examine this aspect here as the main focus of this article is rather to highlight a systematic weakness in the existing approach to XAI and to point out what is necessary to fix it.

Another limitation, that is shared by many other works on causality, is the assumption that the causal graph that has generated the data is known. Once we have obtained this graph, we can often fit the functions $(f_X)_{X\in\mathcal{V}}$ in the SCM and subsequently apply Algorithm 1. This is the procedure that we follow for the real world example in Section 4 below. In specific cases, such as linear SCMs with non-Gaussian noise, algorithms such as LiNGAM [Shimizu et al., 2011, Tashiro et al., 2014, Leyder et al., 2025, Zhang et al., 2025] can be used, cf. Section 4 below, to obtain the causal graph. For more general approaches to causal discovery compare also, e.g., the works Zheng et al. [2018], Bello et al. [2022], Pamfil et al. [2020] and the reviews by Zanga et al. [2022], Glymour et al. [2019]. In the general case, however, obtaining a valid causal graph is typically non-trivial and often requires expert knowledge. Moreover, in accordance with many other works in causal inference, we have to assume that the exogenous variables in our model are uncorrelated, which is typically an oversimplification.

The use of a causal graph further implies that a variable has a specific static meaning throughout the dataset, which is typically not true for other types of data such as images. Here causal representation learning becomes necessary to apply causal techniques [Schölkopf et al., 2021, Brehmer et al., 2022, Ahuja et al., 2023]. We did not use such techniques here and only restrict ourselves to static data.

Finally, a recent series of works, e.g. [Fumagalli et al., 2023, Muschalik et al., 2024a,b, Pelegrina et al., 2026, Kolpaczki et al., 2026], puts a focus on feature interaction, building on extensions of (1) [Grabisch and Roubens, 1999], rather than on single feature attribution. Although we believe that cc-Shapley values could, in principle, be extended to study interactions, we did not study such aspects in this work.

4 EXPERIMENTAL RESULTS

Linear SCMs. We randomly sample 3,000 linear SCMs with 8 features, i.e., $\mathcal{F} = \{X_1, \dots, X_8\}$, $\mathcal{V} = \mathcal{F} \cup \{Y\}$, and additive non-Gaussian (Laplacian) noise as described in Appendix E.2. For each SCM, the effect of including univariate context $\mathcal{S} = \{X_2\}$ on the importance of the variable X_1 is studied. Using linear models for the fitting of conditional expectations and for Algorithm 1, we obtain partial regression coefficients $b_{X_1|X_2}, b_{X_1|\mathrm{do}(X_2)}, b_{X_1}$ for the X_1-dependency of $I_{X_2}(X_1)$, $I_{\mathrm{do}(X_2)}(X_1)$ and $I_\emptyset(X_1)$.

Figure 2 compares $b_{X_1|X_2}$ vs. b_{X_1} (left) and $b_{X_1|\mathrm{do}(X_2)}$ vs. b_{X_1} (right) for all sampled SCMs. Each point represents a different linear SCM. For the computation of $b_{X_1|\mathrm{do}(X_2)}$

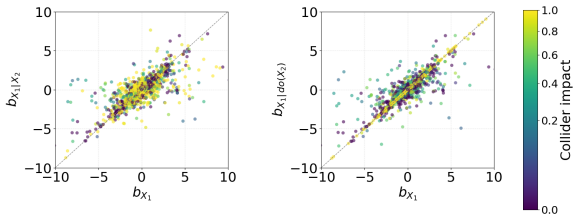


Figure 2: Comparison of the regression coefficients $b_{X_1|X_2}$ (left) and $b_{X_1|\mathrm{do}(X_2)}$ (right) with b_{X_1} for randomly sampled linear SCMs with 9 variables. The color encodes the extent to which X_2 acts as collider, cf. Appendix E.2.

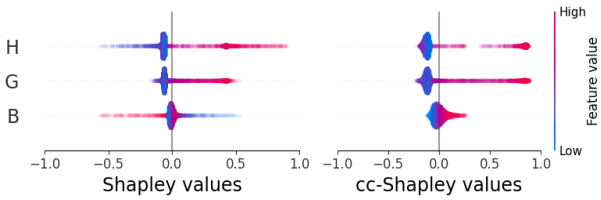


Figure 3: Shapley (left) and cc-Shapley (right) values for the nonlinear diabetes example considered in Section 4.

we used the DirectLiNGAM algorithm from Shimizu et al. [2011] to identify the causal graph, so that no prior knowledge about the causal structure was used. The color of each point represents how many of the paths between X_1 and Y that run through X_2 are blocked due to X_2 acting as a collider. This is measured by a heuristic quantity described in Appendix E.2.

For yellow points, almost all paths between X_1 and Y running through X_2 contain X_2 as a collider, so that we expect the collider bias to be most distinct for these points. We observe that, in the comparison $b_{X_1|X_2}$ vs. b_{X_1} (left), these points are usually placed far away from the diagonal, indicating that including the observed value of X_2 did change the importance of X_1. For the interventional comparison $b_{X_1|\mathrm{do}(X_2)}$ vs. b_{X_1} (right), these points lie on the diagonal and hence context X_2 that mostly affects X_1 via collider bias does not lead to a change in importance attributed to X_1.

Nonlinear Case. Consider again the task of predicting whether a patient has type 2 diabetes ($Y = 1$) or not ($Y = 0$). We assume this time that the patient did obey and not eat breakfast, but measure in addition to the blood glucose G also the average blood sugar H and the BMI B of the patient. The causal graph for this example is shown in Figure 5a, the (synthetic) SCM is specified in Appendix E.3.

It is widely accepted that a high BMI B is associated with an increased risk of type 2 diabetes [Chandrasekaran and Weiskirchen, 2024]. When looking at the Shapley values (1) shown on the left in Figure 3, we might come to a different conclusion: The BMI B is attributed a negative relevance

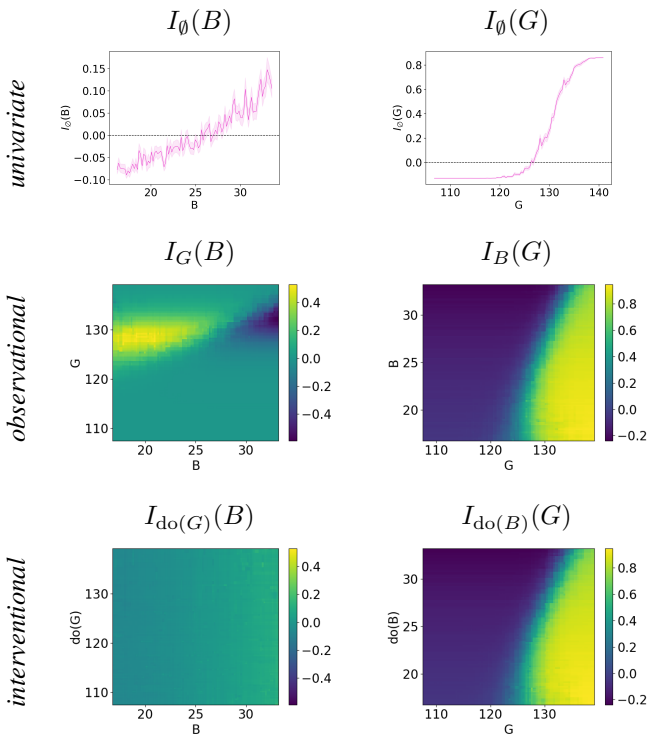
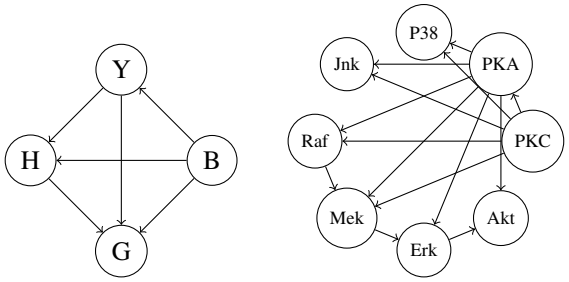


Figure 4: Various terms arising in the computation of the conventional observational Shapley values (1) and the cc-Shapley values (6) for the (more complex) diabetes example introduced in Section 4. The upper row shows (with standard error) the context-free univariate term that coincides for (1) and (6).

for diabetes, which one could easily misinterpret as "high BMI leads to a low diabetes risk". Using cc-Shapley values (6) instead leads to the values depicted on the right in Figure 3. The importance behavior of B (and the other features) now matches intuition. This different behavior of Shapley and cc-Shapley values arises, once more, due to collider bias.

The Shapley values for B in this example are computed as $\phi(B) = \frac{1}{3}I_\emptyset(B) + \frac{1}{6}I_G(B) + \frac{1}{6}I_H(B) + \frac{1}{3}I_{\{G,H\}}(B)$. As shown in the first plot in the top row of Figure 4, the context-free quantity $I_\emptyset(B)$ alone indicates a positive relevance for diabetes risk in line with our expectations. The overall negative relevance of large B must therefore stem from the terms with context. The left plot in the middle row of Figure 4 shows the values of $I_G(B)$ for various choices of B and G as a heatmap. For G in a certain range $I_G(B)$ indicates a negative relevance of B: $I_G(B)$ decreases with increasing B for G roughly between 125 and 135. A similar behavior can be observed for H, cf. Appendix E.3. A look at the causal graph in Figure 5a reveals that H and G act as colliders between B and Y so that conditioning on G, H or $\{G, H\}$ will create a spurious (here negative) association with Y leading to the pattern observed in Figure 3.



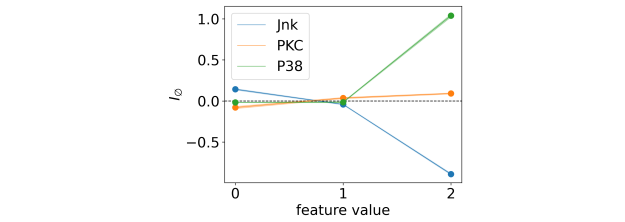
(a) Diabetes and BMI (b) Protein data

Figure 5: Causal graphs for two datasets used in Section 4. For the diabetes example (left) the meaning of the variables are: blood glucose G, average sugar H, BMI B and presence of type 2 diabetes Y (target). For the example from Sachs et al. [2005] (right) the variables are various proteins with known causal graph. The protein *PKA* is here chosen as target.

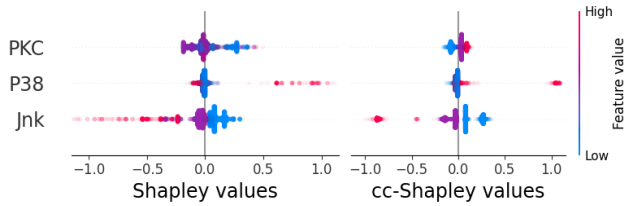
Once we use $I_{\mathrm{do}(G)}(B)$ instead of $I_G(B)$, this bias is removed and we observe no negative relevance of B in the context of G, as shown by the first plot in the third row of Figure 4. Interestingly, when we switch the roles of B and G, we obtain a different effect: Lemma 3.4 implies that $I_{\mathrm{do}(B)}(G) = I_B(G)$ as there are no backdoor paths to B. Both observational and interventional objects now match, cf. the second column of Figure 4, and reveal a positive relevance of G. The reason for the context dependency of $I_{\mathrm{do}(B)}(G)$ is that the relation $B \to G$ is causal and knowing B allows us to remove the variance in G and thus judge the relation between G and Y with more precision. This behavior resembles the observations of Section 3: The suppression effect of B on G is now only accounted for in the relevance of the suppressed variable G and does not affect the importance of the suppressor B. In fact, we have $\phi_{cc}(B) = I_\emptyset(B)$ as we show in Appendix E.3.

A Real World Example. We consider the dataset from Sachs et al. [2005], which contains information on the concentration of various proteins. We use a preprocessed version that discretizes the concentration of each protein in three categories, which we label as $0, 1$ and 2. Details on the dataset and our implementation are contained in Appendix E.4. The causal graph, depicted in Figure 5b, shows the causal relation of the 8 considered proteins. We here consider the task of predicting the amount of the protein *PKA*.

Figure 6a shows the univariate importance $I_\emptyset$ of the three proteins *Jnk, PKC* and *P38* depending on their value in the set {0,1,2}. The results for all features are contained in Appendix E.4. As we saw above in Example 2.1, univariate importance is in general not sufficient to fully explain relevance. However, we also saw above that we have to beware of collider bias when a fundamental deviation in the rele-



(a) Univariate importance $I_\emptyset$



(b) Shapley and cc-Shapley values

Figure 6: Some results of the univariate important measure $I_\emptyset$ and the Shapley and cc-Shapley values for the data from Sachs et al. [2005]. Only a subset of the proteins shown in the causal graph 5b are depicted here. The results for all proteins are contained in Appendix E.4.

vance behavior only occurs in the observational context of other variables.

Figure 6b shows the Shapley values (left) and cc-Shapley values (right) for the proteins *Jnk, PKC* and *P38*. For *Jnk* both, Shapley and cc-Shapley values, show a behavior that is comparable to the univariate importance shown in Figure 6a. For *PKC* and *P38* the Shapley values on the left show a deviation to the univariate behavior. Various data points with a high value of *P38* are attributed negative Shapley values. For *PKC* we see a mixed or even negative relevance in contrast to the slight positive relevance in Figure 6a. Inspecting the causal graph in Figure 5b we see that *PKC* (and hence *P38*) is connected with the target *PKA* via various colliders. Using them as observational context as in (1) therefore bears the risk of introducing collider bias. Indeed, once we use the context only in a causal manner, we find a relevance attribution that contrasts less with the univariate analysis. In particular the (slight) positive relevance of *PKC* is kept intact by the cc-Shapley values.

5 CONCLUSION

In this work, we highlight a blind spot in purely observational approaches to XAI. When studying the relevance of a feature in the observational context of the remaining features, collider bias can distort the feature attribution and even flip the positive relevance of a feature to a negative one and vice versa. This phenomenon can already be observed for hand-crafted two-dimensional problems but also appears for more complex and real world data, and adds to further known biases of XAI methods such as biases towards salient features [Clark et al., 2026]. An incorrect feature attribution can easily lead to misinterpretations and thereby undermine the typical goals pursued when employing XAI.

Eradicating collider bias requires intervention and is hence beyond the reach of the observational rung of Pearl's ladder of causation [Pearl and Mackenzie, 2018] on which most XAI methods fall. Given causal knowledge, we propose cc-Shapley values as an approach to correct for spurious associations introduced by collider bias. We observe in theoretical and experimental results that incorrect and misleading attribution due to collider bias are indeed removed when using cc-Shapley values.

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023.

Christopher J Anders, Leander Weber, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Information Fusion*, 77:261–295, 2022.

Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

Sander Beckers. Causal explanations and xai. In *Conference on causal learning and reasoning*, pages 90–109. PMLR, 2022.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

bnlearn. Reproducing the causal signalling network in Sachs et al., Science (2005). `https://www.bnlearn.com/research/sachs05/`, 2022. Last accessed: 2026-02-10.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35: 38319–38331, 2022.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*, pages 2357–2369. PMLR, 2022.

Gianluca Carloni, Andrea Berti, and Sara Colantonio. The role of causality in explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(2):e70015, 2025.

Preethi Chandrasekaran and Ralf Weiskirchen. The role of obesity in type 2 diabetes mellitus—an overview. *International journal of molecular sciences*, 25(3):1882, 2024.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL `https://doi.org/10.1145/2939672.2939785`.

Benedict Clark, Rick Wilming, Hjalmar Schulz, Rustam Zhumagambetov, Danny Panknin, and Stefan Haufe. Correcting misinterpretations of additive models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Benedict Clark, Marta Oliveira, Rick Wilming, and Stefan Haufe. Feature salience – not task-informativeness – drives machine learning model explanations. *arXiv preprint arXiv:2602.09238*, 2026.

Anthony J Conger. A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and psychological measurement*, 34(1):35–46, 1974.

Richard B Darlington. Multiple regression in psychological research and practice. *Psychological bulletin*, 69(3):161, 1968.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! fundamental misconceptions in current xai research. In *World conference on explainable artificial intelligence*, pages 48–65. Springer, 2023.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33:1229–1239, 2020.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Shap-iq: Unified approximation of any-order shapley interactions. *Advances in Neural Information Processing Systems*, 36: 11515–11551, 2023.

Anders Gjølbye, Stefan Haufe, and Lars Kai Hansen. Minimizing false-positive attributions in explanations of non-linear models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Aric Hagberg, Pieter J Swart, and Daniel A Schult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), 2007.

Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87: 96–110, 2014.

Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Ahcène Boubekki, Jörg Martin, and Danny Panknin. Explainable ai needs formalization. 2024. arXiv preprint arXiv:2409.14590.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33: 4778–4789, 2020.

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9 (4):e1312, 2019.

Paul Horst et al. The role of predictor variables which are independent of the criterion. *Social Science Research Council*, 48(4):431–436, 1941.

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.

Dominik Janzing, Patrick Blöbaum, Atalanti A Mastakouri, Philipp M Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2024.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pages 10476–10501. PMLR, 2022.

Amir-Hossein Karimi. Position: Explainable ai is causal discovery in disguise. 2025.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33: 265–277, 2020.

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.

Yongnam Kim. The causal structure of suppressor variables. *Journal of Educational and Behavioral Statistics*, 44(4): 367–389, 2019.

Patrick Kolpaczki, Felix Edelmann, Maximilian Muschalik, and Eyke Hüllermeier. Approximating shapley interactions with marginal contributions. *Machine Learning*, 115(6):129, 2026.

Gunnar König, Timo Freiesleben, and Moritz Grosse-Wentrup. Improvement-focused causal recourse (icr). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11847–11855, 2023.

Gunnar König, Eric Günther, and Ulrike von Luxburg. Disentangling interactions and dependencies in feature attributions. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2134–2142. PMLR, 03–05 May 2025. URL `https://proceedings.mlr.press/v258/konig25a.html`.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Sarah Leyder, Jakob Raymaekers, and Tim Verdonck. Tslingam: Directlingam under heavy tails. *Journal of Computational and Graphical Statistics*, 34(2):437–447, 2025.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17(4):319–330, 2001.

M.M. Loeve and M.H. Stone. *Probability Theory: University Series in Higher Mathematics*. Literary Licensing, LLC, 2013. ISBN 9781258664541. URL `https://books.google.de/books?id=mUqqmwEACAAJ`.

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

Melkamu Mersha, Khang Lam, Joseph Wood, Ali K Alshami, and Jugal Kalita. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 599:128111, 2024.

Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 55(5):3503–3568, 2022.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. *Advances in Neural Information Processing Systems*, 37: 130324–130357, 2024a.

Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. Beyond treeshap: Efficient computation of any-order shapley interactions for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14388–14396, 2024b.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pages 1595–1605. Pmlr, 2020.

Álvaro Parafita, Tomas Garriga, Axel Brando, and Francisco J Cazorla. Practical do-shapley explanations with estimand-agnostic causal inference. *arXiv preprint arXiv:2509.20211*, 2025.

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

Judea Pearl. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

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

Guilherme Dean Pelegrina, Patrick Kolpaczki, and Eyke Hüllermeier. Shapley value approximation based on k-additive games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17197–17205, 2026.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

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.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005. doi: 10.1126/science.1105809. URL `https://www.science.org/doi/abs/10.1126/science.1105809`.

Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22. Springer, 2019.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Lloyd S Shapley et al. A value for n-person games. 1953.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. Parcelingam: A causal ordering method robust against latent confounders. *Neural computation*, 26(1):57–83, 2014.

Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage*, 110:48–59, 2015.

Rick Wilming, Céline Budding, Klaus-Robert Müller, and Stefan Haufe. Scrutinizing xai using linear ground-truth data with suppressor variables. *Machine learning*, 111 (5):1903–1923, 2022.

Rick Wilming, Leo Kieslich, Benedict Clark, and Stefan Haufe. Theoretical behavior of xai methods in the presence of suppressor variables. In *International Conference on Machine Learning*, pages 37091–37107. PMLR, 2023.

R. Teal Witter, Álvaro Parafita, Tomas Garriga, Maximilian Muschalik, Fabian Fumagalli, Axel Brando, and Lucas Rosenblatt. Exactly computing do-shapley values. *arXiv e-prints*, pages arXiv–2602, 2026.

Felix Wong, Erica J Zheng, Jacqueline A Valeri, Nina M Donghia, Melis N Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, 2024.

Sewall Wright. The method of path coefficients. *The annals of mathematical statistics*, 5(3):161–215, 1934.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Chenglin Zhang, Hong Yu, Guoyin Wang, and Yongfang Xie. Lingam-sf: Causal structural learning method with linear non-gaussian acyclic models for streaming features. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10693–10706, 2025. doi: 10.1109/TNNLS.2025.3529753.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

cc-Shapley: Measuring Multivariate Feature Importance Needs Causal Context (Supplementary Material)

Jörg Martin[1] **Stefan Haufe**[1,2,3]

[1]Physikalisch-Technische Bundesanstalt, Berlin, Germany
[2]Technische Universität, Berlin, Germany
[3]Charité Universitätsmedizin, Berlin, Germany

A ON THE USED VERSION OF SHAPLEY VALUES

Comparison With The Literature. In the context of game theory, Shapley et al. [1953] introduced a general expression of the form

$$\phi(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \gamma(\mathcal{S})(v(\mathcal{S}\cup\{X_j\}) - v(\mathcal{S}))\,, \tag{7}$$

where $\gamma(\mathcal{S}) = \frac{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}{|\mathcal{F}|!}$ and $v : 2^{\mathcal{F}} \to \mathbb{R}$ is a general set function that satisfies some simple properties and should be thought of as encoding a game. Lipovetsky and Conklin [2001] use R^2 for v in the context of linear regression to judge feature importance. Datta et al. [2016] consider a more general framework that uses (7) to measure the influence of inputs with various choices of v. Referring to the works above, Lundberg and Lee [2017] formulate a "classic" version of Shapley values as

$$\phi(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \gamma(\mathcal{S})(f(\mathcal{S}\cup\{X_j\}) - f(\mathcal{S}))\,, \tag{8}$$

here $f(\mathcal{S}\cup\{X_j\})$ and $f(\mathcal{S})$ denote functions that were trained only on the features $\mathcal{S}\cup\{X_j\}$ and $\mathcal{S}$ respectively. The optimally trained models f have precise mathematical formulations in terms of a conditional expectation, cf. Lemma C.1 below. We can thus identify $f(\mathcal{S}\cup\{X_j\}) = \mathbb{E}[Y|\mathcal{S}, X_j]$ and $f(\mathcal{S}) = \mathbb{E}[Y|\mathcal{S}]$ and arrive at the expression (1) used in this work.

Expression (1) is model agnostic in the sense that it only uses optimal models and not a specific instance of a trained model. While Lundberg and Lee [2017] refer to classical Shapley values as formulated in (8), their proposed algorithm SHAP completely builds on the full model $f(\mathcal{F})$. Janzing et al. [2020] explicitly use the output of the model, called $\hat{Y}$ in the following, for their arguments and their algorithm. The approach presented in this work is also usable for $\hat{Y}$ instead of Y. A simplified, pragmatic adaption might be to treat $\hat{Y}$ as an estimate of Y, which is likely to yield comparable results for models with high accuracy (or low MSE in the case of regression). A more causal approach would be to add a node $\hat{Y}$ to the causal graph and draw arrows from each feature used for training to $\hat{Y}$ with the corresponding assignment function $f_{\hat{Y}}$ now just given by the learned machine learning model. Doing so will however not cure the problems inherent to (1) that we outlined in the main text: Consider again the example with the causal diagram 5a. First, note that Y can actually not simply be removed from the diagram as it acts as a mediator (a chain on a causal path) between B and the nodes G and H. Second, there is a collider bias between H and B when conditioning on G so that even with Y completely replaced by $\hat{Y}$ we would have terms in (1) that are influenced by collider bias.

Summarizing, we believe that the model-agnostic approach presented in this work is the most suitable and expressive to highlight feature relevance, as it uses the actual process underlying the data and simplifies matters by ignoring model specific aspects.

On the Combinatorial Coefficients $\gamma(\mathcal{S})$. We mentioned above that there are various versions of Shapley values. These versions only differ in their choice of v in (1) but keep the same coefficients $\gamma(\mathcal{S})$. We here shortly recall the meaning of

$\gamma(\mathcal{S})$ [Osborne and Rubinstein, 1994] to motivate why we also use them in the cc-Shapley values from Definition 3.1:

If we do not want to make any additional assumptions, it is unclear which features $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ to use as a context to judge the relevance of X_j. To decide this in a fair manner, imagine that we base our decision on a randomized experiment: We shuffle the order of features $\mathcal{F} = \{X_1, \dots, X_n\}$ by uniformly drawing from all permutations on $\mathcal{F}$. We then use all variables that come before X_j in such a ordering as context. The probability that the set $\mathcal{S} \subseteq \mathcal{F}$ is chosen as context in such an experiment is then just

$$\gamma(\mathcal{S}) = \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!}\,. \tag{9}$$

Here, $|\mathcal{S}|!$ are the number of ways we can reorder $\mathcal{S}$ in front of X_j, $(|\mathcal{F}| - |\mathcal{S}| - 1)!$ are the number of ways we can order the remaining features after X_j, and the denominator $|\mathcal{F}|!$ is just the number of permutations in $\mathcal{F}$. In particular, we see that the coefficients in (7) add up to 1.

The same logic can be applied to motivate the usage of $\gamma(\mathcal{S})$ for the cc-Shapley values in Definition 3.1: We choose a random subset $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ as causal context using the same randomized procedure.

On the Asymmetry of cc-Shapley Values. In Definition 3.1 we introduce the cc-Shapley values as

$$\phi_{cc}(X_j) = \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}} \gamma(\mathcal{S}) \left(\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(\mathcal{S})]\right). \tag{10}$$

We motivated above the usage of the same coefficients $\gamma(\mathcal{S})$. However, in one other fundamental aspect this definition differs from (7): The objects $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$ and $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$ cannot be understood as set functions as they treat X_j and $\mathcal{S}$ differently. cc-Shapley values therefore lack the symmetry in the treatment of all considered features and, in general, will *not* have the so-called "efficiency" property

$$\sum_j \phi(X_j) = v(\mathcal{F}) - v(\emptyset) \tag{11}$$

satisfied by (7) [Shapley et al., 1953]. In this aspect, this work differs from other works such as [Lundberg and Lee, 2017, Jung et al., 2022, Heskes et al., 2020], which all assign axioms such as (11) a central role for motivating their approaches. We here argue that axioms such as (11) need to be abandoned to provide an XAI method that works in an anti-causal setups and is free from spurious associations. Frye et al. [2020] also consider the classical axioms of Shapley values as too restrictive for XAI purposes and advocate abandoning the symmetry property (but not (11)) to allow for the incorporation of causal knowledge. We compare our work with theirs in Section B below.

From a game theoretical viewpoint our asymmetric modification means that we give the "player" X_j an advantage that we don't grant the other "players" $\mathcal{S}$, namely to use non-causal association. This asymmetric treatment is vital to the performance of ϕ_{cc}. In fact,

- We need to allow for non-causal association between X_j and Y to obtain non-trivial importance in anti-causal setups, cf. Section 3.
- The context variables $\mathcal{S}$ cannot be treated in the same way, as we would otherwise allow X_j to be attributed importance that solely arises from suppressing noise in the variables $\mathcal{S}$ as we observe for Example 1.1.

Breaking the symmetric, equal treatment of all features in (7) is therefore precisely what allows (10) to correct spurious associations.

B DETAILED COMPARISON WITH OTHER METHODS

We here compare cc-Shapley with different methods from the literature that also consider Shapley values in the light of the causal structure of the data. Most of the methods listed below are not designed for an *anti-causal* setting, that is a setting where Y is a cause and not the effect of its features. Suppression as a phenomenon related to collider bias is, however, more naturally linked to such a setting. For the anti-causal setting of Example 1.1, we find that the methods listed below do either not attribute relevance to the informative feature G or also attribute relevance to the suppressor C. The only exception to this are (a modification of) asymmetric Shapley values (Section B.4), where we however observe inconsistent behavior between causal and anti-causal settings. For better readability and comparability, we will rewrite all methods in a similar notation to ours.

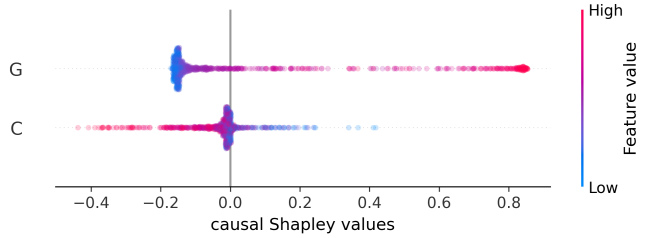


Figure 7: Causal Shapley values as proposed by Heskes et al. [2020] evaluated for the running Example 1.1.

B.1 do-SHAPLEY

Jung et al. [2022] propose do-*Shapley* as a method for measuring causal contributions of each variable, which they define as

$$\phi_{\mathrm{do}}(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \frac{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}{|\mathcal{F}|!}(\mathbb{E}[Y|\mathrm{do}(X_j,\mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(X_j)]\,. \tag{12}$$

Additional aspects regarding the computation of do-Shapley are provided by Parafita et al. [2025] and Witter et al. [2026]. Jung et al. [2022] explicitly require Y to be the last element in the topological order of the graph, whereas Parafita et al. [2025] simply discard all variables that are not ancestors of Y. In fact, since (12) accounts only for causal contributions every non-causal association is discarded. Accordingly, we obtain for Example 1.1

$$\phi_{\mathrm{do}}(C) = \phi_{\mathrm{do}}(G) = 0\,.$$

B.2 CAUSAL SHAPLEY VALUES

Heskes et al. [2020] propose *causal Shapley values* which are given as

$$\phi_{\mathrm{causal}}(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{X_j\}} \frac{|\mathcal{S}|!(|\mathcal{F}|-|\mathcal{S}|-1)!}{|\mathcal{F}|!}(\mathbb{E}[f(\mathcal{F})|\mathrm{do}(X_j,\mathcal{S})] - \mathbb{E}[f(\mathcal{F})|\mathrm{do}(X_j)]\,, \tag{13}$$

with $f(\mathcal{F})$ being the machine learning model under consideration that uses all features $\mathcal{F} = \{X_1,\dots,X_n\}$. The only difference to do-Shapley from Section B.1 is thus the replacement of Y by $f(\mathcal{F})$. For the anti-causal setup of Example 1.1 this creates, however, a fundamental difference in behavior, as $f(\mathcal{F})$ is the effect of all features $\mathcal{F} = \{C, G\}$. Figure 7 shows the values of ϕ_{causal} for $f(\mathcal{F})$ fitted as in Section E.1. For the computation of ϕ_{causal}, we used the known SCM (28) and averaged $f(C, G)$ over 1000 samples from each intervened SCM to compute expectations. We observe that, similar to the behavior of conventional Shapley values, causal Shapley values falsely indicate a negative relevance for the suppressor C.

B.3 INTRINSIC CAUSAL CONTRIBUTION

Janzing et al. [2024] introduce the *intrinsic causal contribution* (ICC) of a feature X_j via

$$\phi_{\mathrm{ICC}}(X_j) = \sum_{\mathcal{S}\subseteq\mathcal{V}\setminus\{X_j\}} \frac{|\mathcal{S}|!(|\mathcal{V}|-|\mathcal{S}|-1)!}{|\mathcal{V}|!}(\psi(Y|U_j,U_\mathcal{S}) - \psi(Y|U_\mathcal{S}))\,, \tag{14}$$

where $\psi(Y|U_\mathcal{S})$ either denotes $\mathbb{E}[\mathrm{Var}(Y|U_\mathcal{S})]$ (for regression) or the conditional entropy $H(Y|U_\mathcal{S})$ (for classification). Note that, in contrast to all other methods discussed in this section, (14) really sums over subsets of $\mathcal{V} = \mathcal{F}\cup\{Y\}$, i.e. subsets that might include Y. The reason behind this is that exogenous noise variables are considered instead of the actual features, an idea that was previously used in Budhathoki et al. [2022] for the purpose of detecting root causes behind outliers.

For Example 1.1, we have for any $X \in \{C, G\}$ and $U_T \subseteq \{U_C, U_G, U_Y\}\setminus\{U_X\}$ that

$$\psi(Y|U_T) - \psi(Y|U_T, U_X) = 0\,, \tag{15}$$

since $Y \perp\!\!\!\perp U_X|U_T$. Thus, similar as for do-Shapley from Section B.1, we find that the ICC framework does not assign importance to any feature in Example 1.1.

(a) Causal mediation

(b) Anti-causal mediation

Figure 8: Simple cases of causal (left) and anti-causal (right) mediation between X and Y through M on which asymmetric Shapley values show inconsistent behavior.

B.4 ASYMMETRIC SHAPLEY VALUES

Frye et al. [2020] introduce *asymmetric Shapley values*, which they define as

$$\phi_{\text{asym.}}(X_j) = \sum_{\pi} w_\pi(\mathbb{E}[f(\mathcal{F})|\{X_i|\pi(i) \leq \pi(j)\}) - \mathbb{E}[f(\mathcal{F})|\{X_i|\pi(i) < \pi(j)\}])\,. \tag{16}$$

Here the sum runs over all permutations π on $\{1,\dots,n\}$ and the $w_\pi \geq 0$ add up to 1. Frye et al. [2020] suggest to restrict $w_\pi \propto 1$ to only those permutations that are coherent with the topological ordering of the graph. In the form (16) discussed by Frye et al. [2020], asymmetric Shapley values are designed to work in a causal setting as $f(\mathcal{F})$ can be considered as the effect of all features, cf. Section B.2. However, if we assume that $f(\mathcal{F})$ is the optimal model $\mathbb{E}[Y|\mathcal{F}]$ (cf. Section C.1), we can actually rephrase (16) as

$$\phi_{\text{asym.}}(X_j) = \sum_{\pi} w_\pi(\mathbb{E}[Y|\{X_i|\pi(i) \leq \pi(j)\}) - \mathbb{E}[Y|\{X_i|\pi(i) < \pi(j)\}])\,, \tag{17}$$

where we used the tower property of conditional expectations to replace $f(\mathcal{F}) = \mathbb{E}[Y|\mathcal{F}]$ with Y within the conditional expectations of (16). In the following, we discuss the behavior of (17) in the presence of suppression and in anti-causal settings.

For Example 1.1 there is only one permutation π that respects the topological ordering of $\mathcal{F} = \{G, C\}$ and which yields

$$\begin{aligned}\phi_{\text{asym.}}(C) &= \mathbb{E}[Y|C] - \mathbb{E}[Y] = 0\,,\\ \phi_{\text{asym.}}(G) &= \mathbb{E}[Y|G,C] - \mathbb{E}[Y|C] = I_C(G) \neq 0\,.\end{aligned} \tag{18}$$

Thus, asymmetric Shapley values (16) correctly indicate relevance of G and no relevance for C. As can be seen in (18), they can however enforce the usage of context: $\phi_{\text{asym.}}(G)$ has no term comparable to $I_\emptyset$ from (1). This can lead to an inconsistent behavior in anti-causal settings when compared with a causal setting. To see this, consider the two simple cases illustrated in Figure 8. Figure 8a show a scenario where X influences Y through a mediator M. The asymmetric Shapley value for X is given as

$$\phi_{\text{asym.}}(X) = \mathbb{E}[Y|X] - \mathbb{E}[Y] \tag{19}$$

which will, in general, be different from zero and thus X is marked as relevant. When using the anti-causal meditation scenario in Figure 8b, however, we obtain

$$\phi_{\text{asym.}}(X) = \mathbb{E}[Y|X,M] - \mathbb{E}[Y|M] = 0\,, \tag{20}$$

since $X \perp\!\!\!\perp Y|M$. Once more, as for Example 1.1, we are forced to use context which yields in this case zero relevance of X. In other words, the impact of mediation on the relevance of a feature fundamentally differs between causal and anti-causal settings.

In summary, asymmetric Shapley values in the modified form (17) are the only method discussed in this section which is well-behaved on Example 1.1. However, in contrast to their behavior in a causal setting, and to the cc-Shapley values from Definition 3.1(which always possess a no-context term $I_\emptyset(X)$) , they are sensitive to mediation in an anti-causal setting.

C FURTHER THEORETICAL ASPECTS

C.1 CONDITIONAL EXPECTATION AS OPTIMAL MODEL

The following basic result recalls that the mathematically optimal model for a typical (supervised) machine learning problem is given by the conditional expectation. In particular, this shows that we can obtain an estimate of conditional expectations

such as $\mathbb{E}[Y|\mathcal{S}]$ with features $\mathcal{S} \subseteq \mathcal{F}$ simply by training a data-driven model with input $\mathcal{S}$ and target Y using MSE-loss (for regression) or cross-entropy loss (for classification).

Lemma C.1 (Conditional expectation as optimally trained model)**.** *Consider a set of square-integrable random variables $\mathcal{S}$ and another square integrable and real-valued random variable Y defined on the same probability space. Then the optimization problem*

$$\arg\min_f \mathbb{E}[|Y - f(\mathcal{S})|^2] \tag{21}$$

has (almost surely) a unique solution given by the conditional expectation $\mathbb{E}[Y|\mathcal{S}]$. If Y is binary, then $\mathbb{E}[Y|\mathcal{S}] = P(Y = 1|\mathcal{S})$ is moreover the optimal solution to the problem

$$\arg\min_f \mathbb{E}[\mathcal{L}_{\mathrm{CE}}(Y; f(\mathcal{S}))]\,, \tag{22}$$

where $\mathcal{L}_{\mathrm{CE}}(Y; f(S)) = -Y \log f(\mathcal{S}) - (1 - Y)(1 - \log f(\mathcal{S}))$ is the cross-entropy loss.

Remark C.2. A similar statement holds for the multivariate case or multiclass case (when Y is one-hot-encoded).

Proof. The statement on $L2$-optimality (21) is a classical result from probability theory [Loeve and Stone, 2013]. The statement based on (22) can be seen by using the tower property of conditional expectations:

$$\begin{aligned}&\mathbb{E}[\mathcal{L}_{\mathrm{CE}}(Y; f(\mathcal{S}))] = \mathbb{E}[\mathbb{E}[\mathcal{L}_{\mathrm{CE}}(Y; f(\mathcal{S}))|\mathcal{S}]]\\ &= -\mathbb{E}[p_{\mathcal{S}} \log f(\mathcal{S}) + (1 - p_{\mathcal{S}})(1 - \log f(\mathcal{S}))]\,,\end{aligned}$$

where $p_{\mathcal{S}} = P(Y = 1|\mathcal{S})$. Minimizing the inner expression w.r.t $f(\mathcal{S})$ (setting the derivative to zero), yields $f(\mathcal{S}) = p_{\mathcal{S}} = P(Y = 1|\mathcal{S}) = \mathbb{E}[Y|\mathcal{S}]$. □

C.2 PROOFS OF LEMMAS 3.3 AND 3.4

We here restate the lemmas from Section 3.2 together with their proofs.

Lemma C.3 (Irrelevant context)**.** *Consider $X_j \in \mathcal{F}, \mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ and assume that there are no causal paths $\mathcal{S} \dashrightarrow Y, X_j$. Then we have $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|X_j]$ and $I_{\mathrm{do}(\mathcal{S})}(X_j) = I_\emptyset(X_j)$.*

Proof. We apply rule 3 of do calculus [Pearl, 2009] to remove $\mathrm{do}(\mathcal{S})$ from the conditional expectation, i.e., we have to verify the following d-separation

$$Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{\mathcal{S}}}} \mathcal{S} \mid X_j\,, \tag{23}$$

where $\mathcal{G}_{\overline{\mathcal{S}}}$ denotes the graph where all incoming edges to $\mathcal{S}$ were deleted. The actual rule 3 uses $\mathcal{G}_{\overline{\mathcal{S}(X_j)}}$ instead of $\mathcal{G}_{\overline{\mathcal{S}}}$, which denotes the removal of all ancestors of X_j from $\mathcal{S}$. By assumption, $\mathcal{S}$ contains, however, no ancestors of X_j. Consider, conditional on X_j, a potential unblocked path between $\mathcal{S}$ and Y in $\mathcal{G}_{\overline{\mathcal{S}}}$. Denote by $X_k \in \mathcal{S}$ the last element within $\mathcal{S}$ on the way to Y and consider, from now on, the unblocked sub-path from X_k to Y. As we removed ingoing edges to $\mathcal{S}$, this path must have an outgoing edge at X_k. By assumption, there is no path $X_k \dashrightarrow Y$. The only remaining path with an outgoing edge at X_k in $\mathcal{G}_{\overline{\mathcal{S}}}$ that is unblocked conditional on X_j can be a path on which X_j or its ancestors $\mathrm{an}(X_j)$ serve as colliders and there are no other colliders on this path. Let $C \in \mathrm{an}(X_j) \cup \{X_j\}$ denote the first of these colliders. We know that there is no other collider between X_k and C and that the path leaves X_k with outgoing edge, we must thus have $X_k \dashrightarrow C$ (the arrows cannot flip direction in between). However, this cannot be true as $C \in \mathrm{an}(X_j) \cup \{X_j\}$ would imply that there is a path $X_k \dashrightarrow X_j$ which violates our assumption. This shows (23) and hence $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|X_j]$.

As we assumed that there are no causal paths $\mathcal{S} \dashrightarrow Y$ we have further $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})] = \mathbb{E}[Y]$ from which the identity $I_{\mathrm{do}(X_k)}(X_j) = I_\emptyset(X_j)$ follows. □

Lemma C.4 (Intervention equals observation)**.** *Consider $X_j \in \mathcal{F}$, $\mathcal{S} \subseteq \mathcal{F} \setminus \{X_j\}$ such that, either*

- (no backdoor paths) *there are no unblocked backdoor paths from $\mathcal{S}$ to X_j or from $\mathcal{S}$ to Y, or*
- (purely causal setup) *there are no causal paths $Y \dashrightarrow X_j, \mathcal{S}$ and no confounders $H \in \mathcal{V} \setminus (\{Y, X_j\} \cup \mathcal{S})$ with $X_j \dashleftarrow H \dashrightarrow Y$ or $\mathcal{S} \dashleftarrow H \dashrightarrow Y$,*

then we have the identities $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|X_j, \mathcal{S}]$ *and* $I_{\mathrm{do}(\mathcal{S})}(X_j) = I_{\mathcal{S}}(X_j)$.

Proof. We will show the identity for the conditional expectation first. For both scenarios, "no backdoor paths" and "purely causal setup" we will use rule 2 of do calculus [Pearl, 2009] to replace the do-operation by conditioning. We have to show

$$Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{\mathcal{S}}}} \mathcal{S}|X_j \,, \tag{24}$$

where $\mathcal{G}_{\underline{\mathcal{S}}}$ denotes the graph where all outgoing edges from $\mathcal{S}$ were removed. Consider first the "no backdoor paths" scenario: Any path between $\mathcal{S}$ and Y in $\mathcal{G}_{\underline{\mathcal{S}}}$ that is unblocked conditional on X_j will have an incoming edge when leaving $\mathcal{S}$ (as it runs in $\mathcal{G}_{\underline{\mathcal{S}}}$) and either 1) not contain X_j or 2) contain X_j or ancestors of X_j as colliders and no other colliders. In the first case, we have just an unblocked backdoor path from $\mathcal{S}$ to Y and in the second case we can subtract an unblocked subpath which is a backdoor path between $\mathcal{S}$ and X_j. Both violate our assumption and therefore we conclude (24).

For the "purely causal setup" scenario consider again a path between $\mathcal{S}$ and Y in $\mathcal{G}_{\underline{\mathcal{S}}}$ that is unblocked conditional on X_j. Denote by $X_k \in \mathcal{S}$ the last element of this path within $\mathcal{S}$ on the way to Y and consider from now on the unblocked subpath between X_k and Y. Since we removed outgoing edges from X_k this path can only have an incoming edge to X_k. Due to our assumptions this path cannot be of the form $X_k \dashleftarrow Y$. The only remaining paths have a fork. This fork cannot be $\leftarrow X_j \rightarrow$ as it would be blocked conditional on X_j. We also know, by assumption, that there is no H such that $X_k \dashleftarrow H \dashrightarrow Y$. The arrows must thus flip directions before or after the fork, which means that there is a collider. The only way this cannot lead to a blocked path is that X_j or a set of ancestors of X_j are these colliders and there are no other colliders. We will write $\mathcal{C} \subseteq \{X_j\} \cup \mathrm{an}(X_j)$ for this set. Call $C \in \mathcal{C}$ the last element in $\mathcal{C}$ on the way of the considered path to Y. Since C is a collider, as there is no further collider between C and Y on the path and as the remaining bit of the path between C and Y is unblocked we must either have $C \dashleftarrow H \dashrightarrow Y$ for some H or $C \dashleftarrow Y$, depending on the direction of the last edge of the path. In either case we obtain due to $C \in \{X_j\} \cup \mathrm{an}(X_j)$ that either $X_j \dashleftarrow H \dashrightarrow Y$ or $X_j \dashleftarrow Y$, which both violate our assumptions. Hence, there is no unblocked path and (24) follows.

We still have to show $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})] = \mathbb{E}[Y|\mathcal{S}]$, which follows again from rule 2 of do calculus once we know

$$Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{\mathcal{S}}}} \mathcal{S} \,.$$

For the "no backdoor paths" scenario this follows from the assumption that there are no backdoor paths from $\mathcal{S}$ to Y. For the "purely causal setup" scenario we can use that any unconditional unblocked backdoor path is either form $\mathcal{S} \dashleftarrow Y$ or $\mathcal{S} \dashleftarrow H \dashrightarrow Y$, which are cases which we both excluded in our assumption. □

C.3 BACKDOOR ADJUSTMENT

The following lemma provides an adaption of the "backdoor adjustment" [Pearl, 2009, Theorem 3.3.2] for the context of this work.

Lemma C.5 (Version of backdoor adjustment)**.** *Consider* $X_j \in \mathcal{F}$, $\mathcal{S}, \mathcal{W} \subseteq \mathcal{F} \setminus \{X_j\}, \mathcal{S} \cap \mathcal{W} = \emptyset$ *such that*

- *There are no causal paths* $\mathcal{S} \dashrightarrow \mathcal{W} \cup \{X_j\}$
- *Both,* $\mathcal{W}$ *and* $\mathcal{W} \cup \{X_j\}$*, block all backdoor paths between* $\mathcal{S}$ *and* Y.

then we have the identity

$$I_{\mathrm{do}(\mathcal{S})}(X_j) = \mathbb{E}_{\mathcal{W}|X_j}[\mathbb{E}[Y|X_j, \mathcal{S}, \mathcal{W}]] - \mathbb{E}_{\mathcal{W}}[\mathbb{E}[Y|\mathcal{S}, \mathcal{W}]] \,.$$

Proof. Recall that $I_{\mathrm{do}(\mathcal{S})}(X_j) = \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] - \mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$. By the rules for conditional expectations, we have

$$\begin{aligned} \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{W}|X_j, \mathrm{do}(\mathcal{S})}[\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S}), \mathcal{W}]] \text{ and} \\ \mathbb{E}[Y|\mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{W}|\mathrm{do}(\mathcal{S})}[\mathbb{E}[Y|\mathrm{do}(\mathcal{S}), \mathcal{W}]] \,. \end{aligned}$$

Since the sets $\mathcal{W}$ and $\mathcal{W} \cup \{X_j\}$ block all backdoor paths between $\mathcal{S}$ and Y we can replace the inner do operation by conditioning (rule 2 of do calculus [Pearl, 2009]) and obtain

$$\begin{aligned} \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{W}|X_j, \mathrm{do}(\mathcal{S})}[\mathbb{E}[Y|X_j, \mathcal{S}, \mathcal{W}]] \text{ and} \\ \mathbb{E}[Y|\mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{W}|\mathrm{do}(\mathcal{S})}[\mathbb{E}[Y|\mathcal{S}, \mathcal{W}]] \,. \end{aligned}$$

To conclude, we remove the do intervention in the outer expectation using $P(X_j, \mathcal{W}|\mathrm{do}(\mathcal{S})) = P(X_j, \mathcal{W})$, which follows from the fact that there are no causal paths $\mathcal{S} \dashrightarrow \mathcal{W} \cup \{X_j\}$ (rule 3 of do calculus [Pearl, 2009]). □

Application of Lemma C.5 requires specific conditions on the relation between X_j and $\mathcal{S}$ as well as a search for a suitable set $\mathcal{W}$. Once found, we have to find a reasonable estimate for the conditional probability $P(\mathcal{W}|X_j)$, which can be a complicated task, especially for higher dimensional $\mathcal{W}$. In the experiments in Section 4, we therefore found a it more practical to first estimate the assignment functions in the SCM followed by the application of Algorithm 1. Identification via more advanced techniques as in Witter et al. [2026] or Jung et al. [2022] could be a feasible alternative but were not studied in the scope of this work.

C.4 FURTHER DETAILS REGARDING ALGORITHM 1

On Stochastic Interventions. Algorithm 1 uses a stochastic intervention $\mathrm{do}(\mathcal{S} \sim q)$. Such an intervention can be defined as follows. Starting from the original SCM $\mathcal{M} = (\mathcal{V}, (f_X)_{X\in\mathcal{V}}, (U_X)_{X\in\mathcal{V}})$ choose a vector of random variables $(\tilde{U}_S)_{S\in\mathcal{S}} \sim q$ that is independent of the exogenous variables in $\mathcal{M}$. Next, define for $S \in \mathcal{S}$ modified assignment functions via

$$\tilde{f}_S(\tilde{U}_S) = \tilde{U}_S\,.$$

The modified SCM $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$ is then simply given by

$$\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)} = (\mathcal{V}, (f_X)_{X\in\mathcal{V}\setminus\mathcal{S}} \cup (\tilde{f}_S)_{S\in\mathcal{S}}, (U_X)_{X\in\mathcal{V}\setminus\mathcal{S}} \cup (\tilde{U}_S)_{S\in\mathcal{S}})\,.$$

This SCM does not exactly fit the definition of a structural causal model we gave in Section 2.1 as the exogenous variables $(\tilde{U}_S)_{S\in\mathcal{S}} \sim q$ are, in general, not independent of each other. For the purposes of this article this is irrelevant, as we only use stochastic interventions in the context of Algorithm 1 and its justification below is still true for dependant $S \in \mathcal{S}$.

Reasoning Behind Algorithm 1 Algorithm 1 implicitly claims that fitting data-driven models in the modified model $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$ yields $\mathbb{E}[Y|\mathrm{do}(\mathcal{S})]$ and $\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]$, where we denote by q the joint marginal of $\mathcal{S}$ within the original SCM $\mathcal{M}$.

We already recalled in Lemma C.1 the established fact that fitting a data driven model with standard L2-loss (regression) or cross-entropy-loss (classification) gives an estimate of the conditional expectation. We are thus left with the claim that

$$\begin{aligned}\mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|X_j, \mathcal{S}]\,,\\ \mathbb{E}[Y|\mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|\mathcal{S}]\end{aligned} \tag{25}$$

to motivate Algorithm 1. In the causal graph of $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$ there are no edges pointing from $\mathcal{V}\setminus\mathcal{S}$ to $\mathcal{S}$ (there might still be confounding within $\mathcal{S}$, cf. the discussion above). In particular, in $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$ there are no backdoor paths from $\mathcal{S}$ to Y, neither blocked nor unblocked. Hence rule 2 of do calculus [Pearl, 2009] applies and we have

$$\begin{aligned}\mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|X_j, \mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|X_j, \mathcal{S}]\,,\\ \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|\mathrm{do}(\mathcal{S})] &= \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|\mathcal{S}]\,.\end{aligned} \tag{26}$$

Now, once we intervene on $\mathcal{S}$ via $\mathrm{do}(\mathcal{S})$ to set it to a fixed value, the actual marginal of q does not influence the distribution of all non-intervened variables $\mathcal{V}\setminus\mathcal{S}$ so that we can perform the transition $\mathcal{M}^{\mathrm{do}(\mathcal{S})} \to \mathcal{M}$ in the following equations

$$\begin{aligned}\mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|X_j, \mathrm{do}(\mathcal{S})] &= \mathbb{E}[Y|X_j, \mathrm{do}(\mathcal{S})]\,,\\ \mathbb{E}_{\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}}[Y|\mathrm{do}(\mathcal{S})] &= \mathbb{E}[Y|\mathrm{do}(\mathcal{S})]\,.\end{aligned} \tag{27}$$

Combing (27) with (26), we obtain the identity (25) that underpins Algorithm 1.

The Exact Choice of q Doesn't Matter. Note, that the arguments above never used the fact that q is the marginal of $\mathcal{S}$ within the original model $\mathcal{M}$. In fact, we can take any distribution q which is supported on all $\mathcal{S}$ values which we want to set as context. This might appear counterintuitive at first glance, but is rooted in the fact that in the modified SCM, intervention and conditioning on $\mathcal{S}$ have the same effect. However, once we intervene via $\mathrm{do}(\mathcal{S} = s)$ the variable $\mathcal{S}$ becomes a constant and the original distribution q of $\mathcal{S}$ in the modified SCM $\mathcal{M}^{\mathrm{do}(\mathcal{S}\sim q)}$ is irrelevant.

Another possible choice for q might be to combine the marginals of $X_k \in \mathcal{S}$ in an independent manner. For the experiments in Section 4 we found that such a proceeding gives indeed practically indistinguishable results to the ones presented here.

On the Implementation of $q(\mathcal{S})$. For the experiments in this article the construction of the sampler for $q(\mathcal{S})$ was done be creating for each $\mathcal{S}$ a separate copy of the SCM whose sole purpose is to sample from $q(\mathcal{S})$. Due to the small size of the problems studied in this article this created no substantial computational overhead. For larger problems, the invariance pointed out above could, however, be exploited to simply replace $q(\mathcal{S})$ by a more feasible alternative (e.g., a uniform distribution with suitable boundaries) to save memory and computational time.

D FURTHER DISCUSSION REGARDING LIMITATIONS.

Computational Complexity Is Bounded by $\mathcal{O}(2^{c|\mathcal{F}|})$. As pointed out in Section 3.3, in the current form, cc-Shapley values are not feasible for higher dimensions. Training models for cc-Shapley requires, in general, 2 steps:

1. Fitting assignment functions for each variable in the causal graph. This results in a SCM $\mathcal{M}$.
2. Application of Algorithm 1 for each $X_j \in \mathcal{F}$ and $\mathcal{S} \subseteq \mathcal{F}$, that is:
   (a) Creating a sampler $q(\mathcal{S})$.
   (b) Sampling of training data.
   (c) Fitting a model on the training data.

As both steps require selecting all $\mathcal{S} \subseteq \mathcal{F}$ and then performing steps of at most polynomial order in $\mathcal{S}$ this leads to a time and memory complexity bounded by $\mathcal{O}(2^{c|\mathcal{F}|})$ for some $c > 1$. cc-Shapley thus scales, similar as classical Shapley values, exponentially in the number of features.

For the main experiments of this work, the total computational time (without GPU usage) is specified in Table 2.

Experiment	Ex. 1.1 (Fig. 1)	Linear SCMs (Fig. 2)	Diabetes and BMI (Fig. 3	Protein (Fig. 6b)
time (total)	1m10s	49m40s	3m21s	17m58s
time (cc-Shapley)	42s	/	2m35s	12m27s

Table 2: Computational time behind the experiments in this work. The last row shows the time needed to produce results only related to cc-Shapley (i.e. without computing classical Shapley values). For the linear SCM experiment, no Shapley or cc-Shapley values were computed and hence the time needed to produce Figure 2 is reported. All experiments were done on CPUs only.

On Hidden Confounders. We assumed in Section 2 that there are no hidden confounders, which is an assumption quite often made in causal inference but one that is, if at all, only approximately true in practice. Generally speaking, this assumption is not necessary as long as all objects appearing in (6) are identifiable. However, as we actually loop through all features and all context sets at the end of the day, full identifiability in a hidden confounder setting will rarely be the case in interesting scenarios.

However, there is an interesting aspect worth mentioning here. Consider once more the example of diabetes and BMI from Figure 5a and suppose that we are not interested in the relevance of all features but only want to study whether the BMI B has a negative or positive relevance for the presence of diabetes Y. In this case a possible hidden confounder $B \leftarrow H \rightarrow Y$ (such as lifestyle) would not hinder the computation of $\phi_{cc}(B)$ as we are not applying the do-Operation on B in (5). Our conclusions can therefore be drawn independently of the possible existence of H.

E EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

Throughout this work, with the exception of the linear SCM experiment, we used `XGBoost` [Chen and Guestrin, 2016] for fitting data-driven models. The exact implementation details can be found in the repository `https://github.com/braindatalab/cc_shapley` that contains the Python source code behind this experimental results in this article.

E.1 DIABETES AND BREAKFAST EXAMPLE (EXAMPLE 1.1)

Example 1.1 from Section 1 can be described by an SCM of the following form

$$\begin{aligned} C &= U_C,\, U_C \sim \mathcal{N}(60; 25^2)\,, \\ Y &= U_Y,\, U_Y \sim \text{Bernoulli}(0.15)\,, \\ G &= 85 + 0.4 \cdot C + 40 \cdot Y + U_G, U_G \sim \mathcal{N}(0; 10^2)\,, \end{aligned} \tag{28}$$

where $\text{Bernoulli}(p)$ denotes the Bernoulli distribution with success probability p and $\mathcal{N}(\mu; \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. For computing the Shapley values (1) displayed in Figure 1b (left), we fitted all conditional expectations via XGBoost on $3 \cdot 10^6$ random samples from (28), using Lemma C.1 above. For the cc-Shapley values (6), we applied Algorithm 1: For each term in (6), we first modified the SCM accordingly and then used, once more, XGBoost on $3 \cdot 10^6$ random samples to fit each conditional expectation. Figure 1b was then produced by evaluating (1) and (6) on 10^4 (new) data points drawn from (28).

All I objects involved in the computation of the Shapley values are plotted in Figure 9. For cc-Shapley the according objects are plotted in Figure 10. The univariate terms $I_\emptyset(C)$ and $I_\emptyset(G)$, shown on the diagonal in Figures 9 and 10, agree for both approaches by definition.

Derivation of $I_C(G)$. In Section 3.1 we gave an explicit expression for $I_C(G)$, which we derive here. By definition we have, using the independence of Y and C,

$$\begin{aligned} I_C(G) &= \mathbb{E}[Y|G,C] - \mathbb{E}[Y|C] \\ &= P(Y=1|G,C) - P(Y=1|C) \\ &= P(Y=1|G,C) - P(Y=1) \\ &= P(Y=1|G,C) - 0.15\,. \end{aligned}$$

Thus, it remains to estimate $P(Y=1|G,C)$. From the SCM (28) we obtain the joint distribution

$$P(Y,G,C) = \text{Bernoulli}(Y|0.15) \cdot \mathcal{N}(C|60; 25^2) \cdot \mathcal{N}(G|0.4C + 85 + 40Y; 10^2)\,.$$

Writing $p = 0.15$ we can use this expression to compute

$$\begin{aligned} P(Y=1|G,C) &= \frac{P(Y=1,G,C)}{P(Y=1,G,C) + P(Y=0,G,C)} \\ &= \frac{1}{1 + \frac{1-p}{p} \frac{\mathcal{N}(G|0.4C+85;10^2)}{\mathcal{N}(G|0.4C+125;10^2)}} \\ &= \frac{1}{1 + \frac{1-p}{p} e^{-\frac{1}{2\cdot 10^2}[(G-0.4C-85)^2 - (G-0.4C-125)^2]}} \\ &= \frac{1}{1 + e^{-[\frac{2}{5}(G-0.4C)-42-\log\frac{1-p}{p}]}} \\ &= \sigma\left(\frac{2}{5}\left(G - 0.4C - 105 - \frac{5}{2}\log\frac{17}{3}\right)\right) \\ &\simeq \sigma\left(\frac{2}{5}(G - 0.4C - 109)\right)\,, \end{aligned}$$

where σ denotes the sigmoid function. In total, we thus have

$$I_C(G) \simeq \sigma\left(\frac{2}{5}(G - 0.4C - 109)\right) - 0.15\,.$$

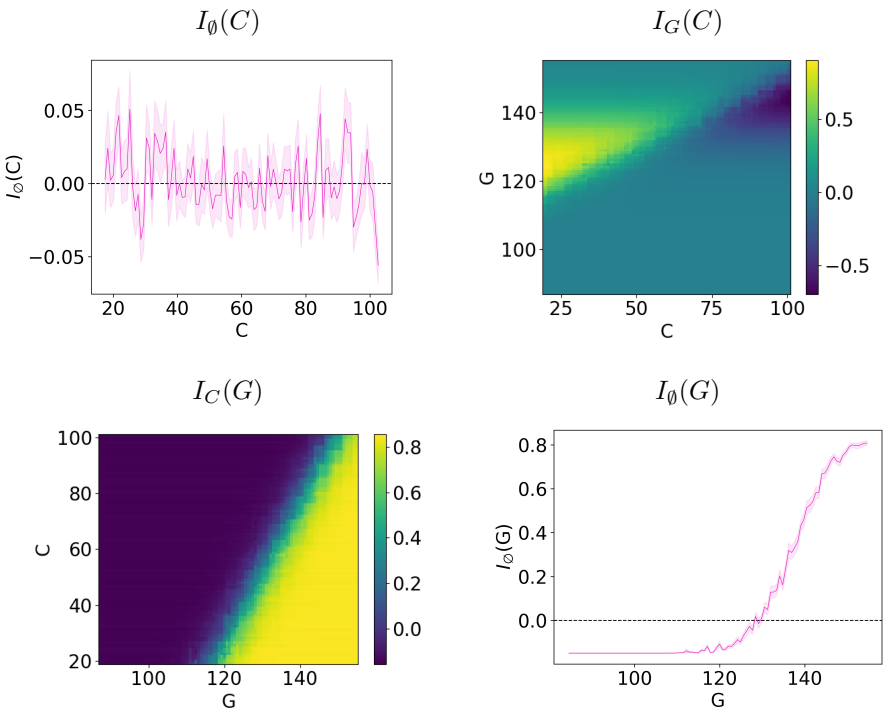


Figure 9: Plot of all objects arising in the computation of the conventional Shapley values (1) for Example 1.1.

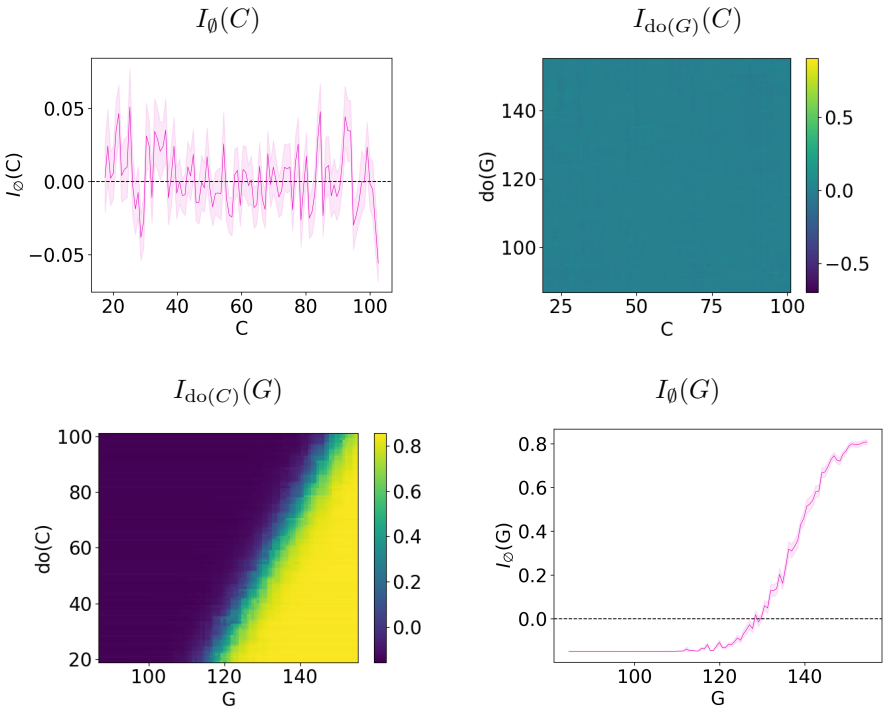


Figure 10: Plot of all objects arising in the computation of the cc-Shapley values (6) for Example 1.1.

E.2 LINEAR SCM EXPERIMENT

For the experiment underlying Figure 2, we sampled 3,000 linear SCMs with 9 variables (including the target). To this end, we applied Algorithm 2 with $n = 9$ and $p = 0.8$ to obtain a random adjacency matrix A and then used i.i.d. random variables $U_1, U_2, \ldots, U_9 \sim \text{Laplace}(0; 0.1)$ (Laplace distribution with mean 0 and scale 0.1) to define the linear SCM

$$\mathbf{V} = A\mathbf{V} + \mathbf{U}\,, \tag{29}$$

where $\mathbf{V} \in \mathbb{R}^9$ denotes the variables of the SCM and $\mathbf{U} = (U_i)_{1 \leq i \leq 9}$ collects the noise variables. For each noise instance $\mathbf{U}$, equation (29) was solved for $\mathbf{V}$ via $\mathbf{V} = (I - A)^{-1}\mathbf{U}$.

Algorithm 2: Sample adjacency matrix of a linear SCM

Input: Number of variables n, edge probability $p \in [0, 1]$
Output: adjacency matrix $A \in \mathbb{R}^{n \times n}$

`# Create random binary matrix`
Sample i.i.d. $U_{i,j} \sim \text{Uniform}(0, 1)$ for $1 \leq i, j \leq n$;
Create $A \leftarrow [U_{i,j} < p]_{1 \leq i,j \leq n}$;

`# Turn into binary adjacency matrix`
For all $i \geq j$ set $A_{ij} \leftarrow 0$;
Draw a random n-permutation π ;
Permute rows and columns of A with π;

`# Fill with random floats`
Sample i.i.d. $\alpha_{ij} \sim \mathcal{N}(0; 1)$ for $1 \leq i, j \leq n$;
Replace $A_{ij} \leftarrow A_{ij} \cdot \alpha_{ij}$;

return A

From each SCM constructed as above, we sampled $3 \cdot 10^4$ samples of $\mathbf{V}$. Target and features are constructed from $\mathbf{V}$ via the assignment:

$$Y = V_1\,, \tag{30}$$

$$X_i = V_{i+1} \text{ for all } 1 \leq i \leq 8. \tag{31}$$

The analysis behind Figure 2 then focuses on the relation between Y, X_1 and X_2.

Computation of $b_{X_1}, b_{X_1|X_2}, b_{X_1,\text{do}(X_2)}$. Since $I_\emptyset(X_1), I_{X_2}(X_1)$ and $I_{\text{do}(X_2)}(X_1)$ only differ from $\mathbb{E}[Y|X_1], \mathbb{E}[Y|X_1, X_2]$ and $\mathbb{E}[Y|X_1, \text{do}(X_2)]$ via X_1-independent terms we used the latter to obtain $b_{X_1}, b_{X_1|X_2}, b_{X_1,\text{do}(X_2)}$. For the observational quantities $b_{X_1}, b_{X_1|X_2}$ we fitted linear univariate and bivariate models with least-squares loss on the $3 \cdot 10^4$ samples described above. For the causal quantity $b_{X_1|\text{do}(X_2)}$, we made a stochastic intervention on X_2 in (29) as described in Algorithm 1 and then conducted another least-squares fit of a bivariate linear model on $3 \cdot 10^4$ samples of the modified model. We used least-squares fitting, instead of maximum likelihood fitting, due to the optimality of conditional expectation under the $L2$-norm, cf. Lemma C.1 above.

Collider Impact. The "Collider impact" indicated in the colorbar of Figure 2 is measured via a heuristic quantity that is defined as follows:

$$\text{Collider impact} = \frac{|CP_{X_2}|}{|CP_{X_2}| + |UP_{X_2}|}\,, \tag{32}$$

where

$$CP_{X_2} = \sum_{X_1, Y \text{ paths containing } X_2 \text{ as the only collider}} \prod A_{ij}\,,$$

$$UP_{X_2} = \sum_{\text{unblocked } X_1, Y \text{ path containing } X_2} \prod A_{ij}$$

denote the sum of products of entries of the adjacency matrix along the paths indicated under the summation symbols. In particular, if all paths between X_1 and Y that run through X_2 and are either unblocked or contain X_2 as the only collider, all fall in the second category, the "Collider impact" will be 1. If they are all unblocked, "Collider Impact" will be 0. Object (32) is (loosely) motivated by Wright's path tracing rules [Wright, 1934, Pearl, 2013]. Note that we ignore in (32), for simplicity, collider bias induced by descendants of colliders.

For the analysis of paths between X_1 and Y, we used the `NetworkX` package from Hagberg et al. [2007].

E.3 DIABETES AND BMI EXAMPLE

The precise SCM for the nonlinear example outlined in Section 4 was entirely hand-crafted and is given by

$$\begin{aligned} B &= U_B,\, U_B \sim \mathcal{N}(25; 5^2) \\ Y &\sim \text{Bernoulli}(\sigma(-2 + 0.1 \cdot (B - 25))) \\ H &= 5 + 10 \cdot Y + 0.01 \cdot B^2 + U_H,\, U_H \sim \mathcal{N}(0; 1) \\ G &= 90 + 20 \cdot Y + 30 \cdot \sigma(-0.5 \cdot (H - 5)) + B + U_G,\, U_G \sim \mathcal{N}(0; 5^2)\,, \end{aligned} \tag{33}$$

where σ denotes the sigmoid function.

The proceeding for (33) coincides with the simpler example of diabetes detection from Section E.1 with the only exception that we now have to consider different and more variables in the computation of (1) and (6). All involved observational I objects with no or univariate context are depicted in Figure 11. The causal analogues for the cc-Shapley values are shown in Figure 12.

Derivation of $\phi_{cc}(B) = I_\emptyset(B)$. The sets $\mathcal{S} = \{G\}$, $\mathcal{S} = \{H\}$ and $\mathcal{S} = \{G, H\}$ all have no causal paths to Y or B, cf. Figure 5a. Thus, applying Lemma 3.3 to each of them we obtain $I_{\text{do}(G)}(B) = I_{\text{do}(H)}(B) = I_{\text{do}(\{G,H\})}(B) = I_\emptyset(B)$. The claimed identity then follows from the fact that the combinatorial coeffiencents add up to 1 (cf. Section A), in detail

$$\phi(B) = \frac{1}{3} I_\emptyset(B) + \frac{1}{6} I_{\text{do}(G)}(B) + \frac{1}{6} I_{\text{do}(H)}(B) + \frac{1}{3} I_{\text{do}(\{G,H\})}(B) = I_\emptyset(B)\,. \tag{34}$$

Robustness Analysis. To study the robustness of cc-Shapley against deviations in the used SCM we introduce $\alpha_{Y\to H}, \alpha_{B\to H}, \alpha_{Y\to G}$ and $\alpha_{H\to G}$ to rephrase the assignment functions of H and G in (33) as

$$\begin{aligned} H &= 5 + \alpha_{Y\to H} \cdot Y + \alpha_{B\to H} \cdot B^2 + U_H,\, U_H \sim \mathcal{N}(0; 1) \\ G &= 90 + \alpha_{Y\to G} \cdot Y + 30 \cdot \sigma(-\alpha_{H\to G} \cdot (H - 5)) + B + U_G,\, U_G \sim \mathcal{N}(0; 5^2)\,, \end{aligned}$$

We then choose relative deviations $\Delta_{\text{ref}}\alpha_{\cdot\to\cdot}$ between -1 and 1 and modify the actual values $\alpha_{\cdot\to\cdot}$ from (33) as

$$\alpha_{\cdot\to\cdot} = \alpha^{\text{true}}_{\cdot\to\cdot} + \Delta_{\text{ref}}\alpha_{\cdot\to\cdot} \cdot \alpha^{\text{true}}_{\cdot\to\cdot}\,. \tag{35}$$

Figure 13 then shows the deviation of ϕ_{cc} normalized by the sum of original relevances, that is

$$\Delta_{\text{norm}}\phi_{cc}(X) = \frac{\phi'_{cc}(X) - \phi_{cc}(X)}{\sum_{\tilde{X}\in\{H,B,G\}} \phi_{cc}(\tilde{X})}\,, \tag{36}$$

against $\Delta_{\text{ref}}\alpha_{\cdot\to\cdot}$, where $X \in \{H, B, G\}$ and where ϕ'_{cc} denotes the cc-Shapley values computed with the modified coefficients (35).

We observe that for most studied coefficients cc-Shapley appears to be robust against deviations in the specification of the SCM. Only for the link $Y \to G$ cc-Shapley shows a high sensitivity that effects the importance of all features.

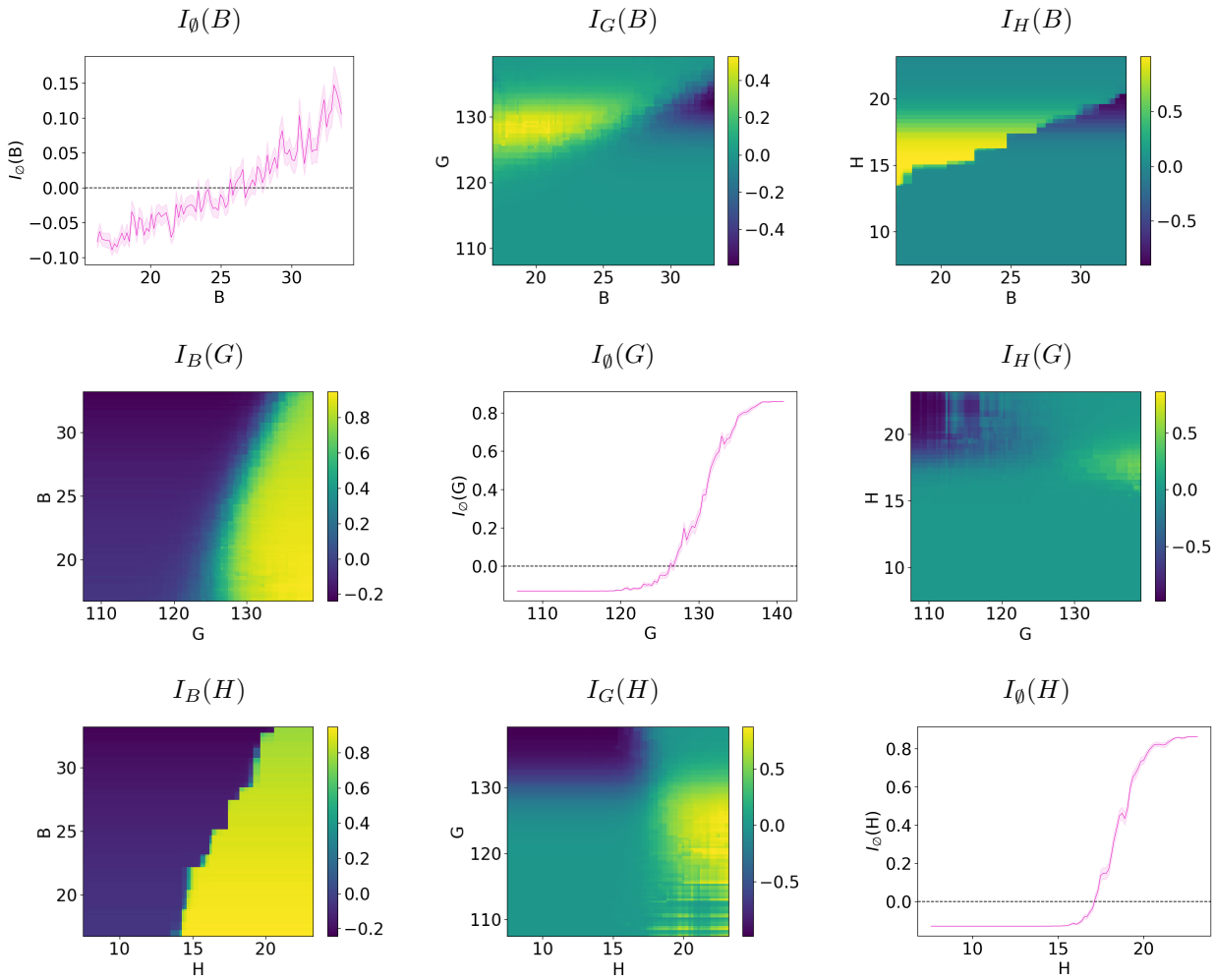


Figure 11: Plot of all objects with no or univariate context arising in the computation of the conventional Shapley values (1) for the SCM (33).

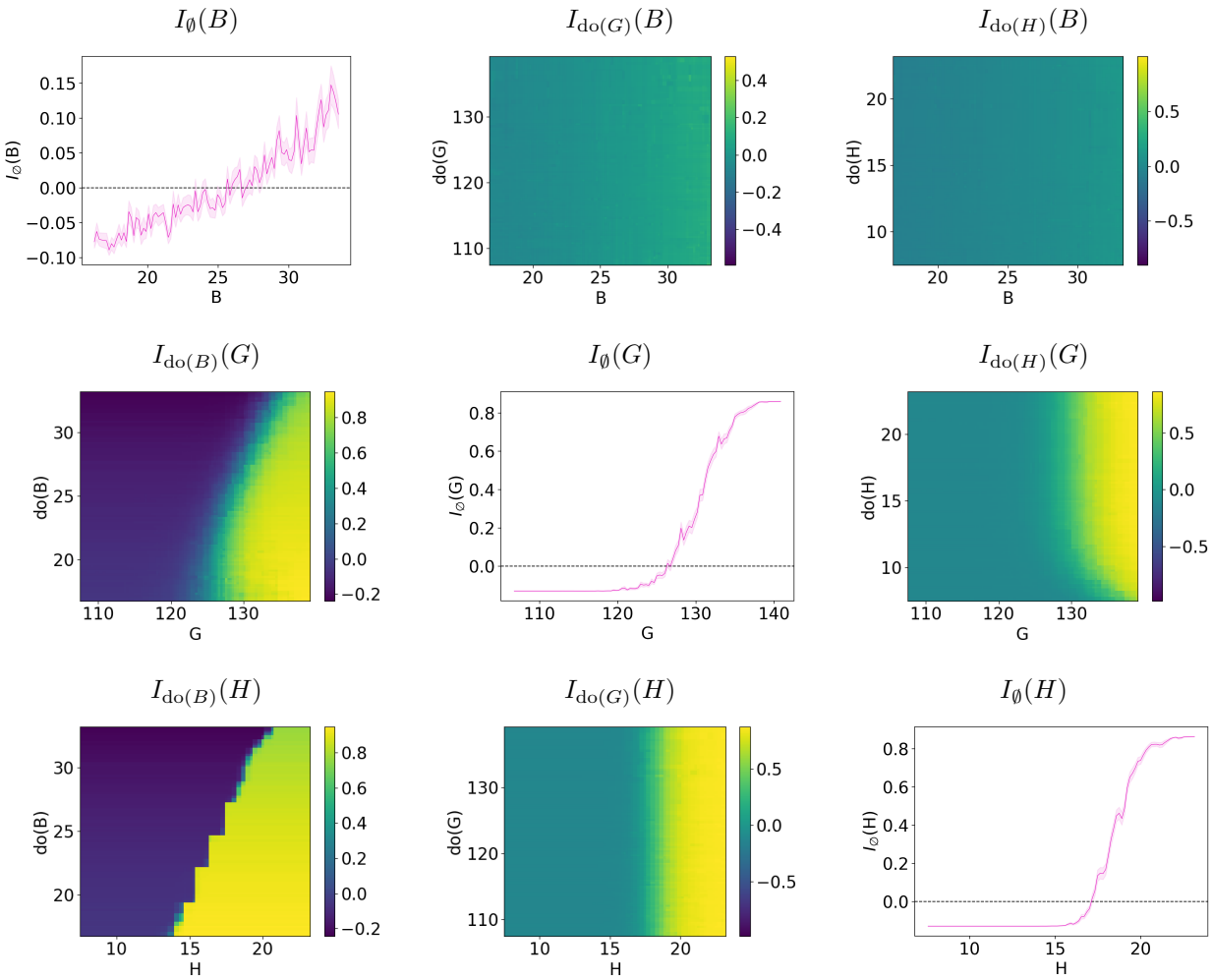


Figure 12: Plot of all objects with no or univariate context arising in the computation of the cc-Shapley values (6) for the SCM (33).

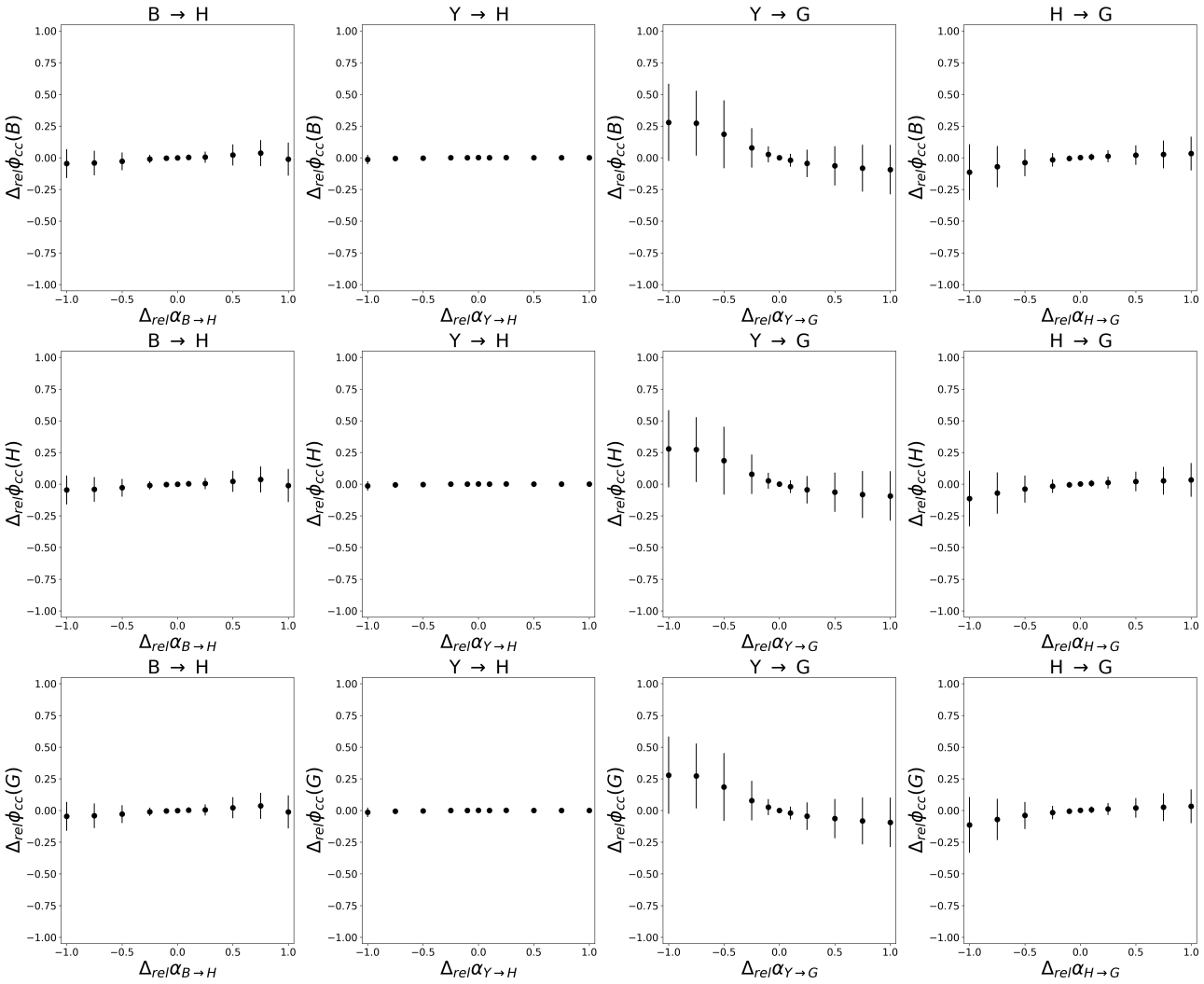


Figure 13: Robustness analysis for the diabetes and BMI example (33). The title on each plot encodes a coefficient $\alpha_{\cdot\to\cdot}$ that is modified, cf. (35). The plots show the normalized deviation $\Delta_{\mathrm{norm}}\phi_{cc}$, cf. (36), of cc-Shapley values against relative deviations of the coefficients $\Delta_{\mathrm{rel}}\alpha_{\cdot\to\cdot}$, cf. (35), ranging from -1 to 1. The markers and error-bars depict the means and standard deviations of $\Delta_{\mathrm{norm}}\phi_{cc}$ evaluated on 10,000 test data points obtained with the original model (33).

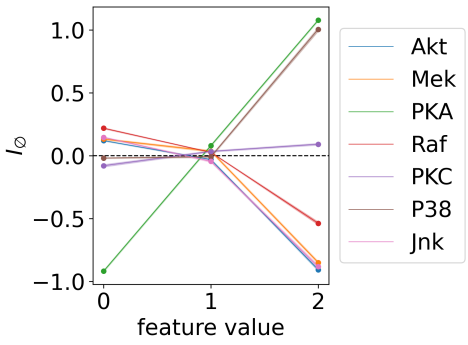
E.4 PROTEIN DATASET FROM Sachs et al. [2005]

We used the preprocessed version of the interventional data from Sachs et al. [2005] available on bnlearn [2022] that bins each variable in the categories $0, 1$ and 2. In addition to the proteins depicted in the causal graph in Figure 5b the data contains further information on three proteins *PIP2,PIP3* and *Plcg*. Sachs et al. [2005] reports their connection to the other proteins in the dataset with a low confidence. We discarded these three proteins for our analysis and only used the ones depicted in Figure 5b. In addition, the data contain an indicator that denotes on which of the proteins an experimental intervention happened. For fitting an SCM, we applied the causal graph from Figure 5b and fitted an XGBoost classifier for each feature using the parents of each feature as input discarding in each case the data points that were obtained with an intervention on the considered feature (if any). Subsequently, these models were used as assignment functions to define an SCM on the variables from Figure 5b. Since each of the assignment function is a classifier, no assumptions on the underlying noise had to be made.

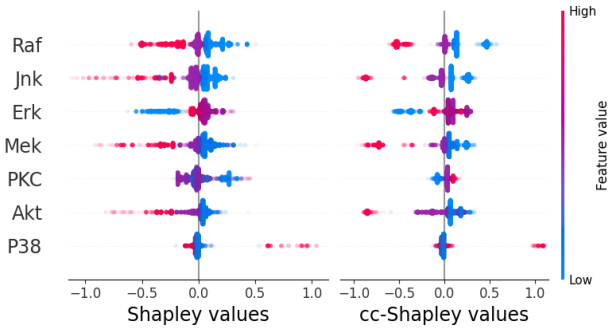
Using the SCM as described above, we then computed the Shapley values 1 and cc-Shapley values (6), training, once more, XGBoost classifiers on 10^4 samples produced via the SCM from above or intervened versions (as described in Algorithm 1). As target Y, we used the binned concentration of the protein *PKA*. Y can take thus the values $0, 1$ and 2 so that the conditional expectations in (1) and (6) average these numbers using the probabilities produced by the learned classifiers. We evaluated (1) and (6) on 10^4 (new) datapoints obtained with the learned SCM.

In Figure 6b in Section 4, we only showed a subset of the Shapley values and cc-Shapley values. The entire set, including all univariate feature importances, is depicted in Figure 14.

Sensitivity of cc-Shapley With Respect to the SCM. Figure 15 shows cc-Shapley values for data generated by slighly modified SCMs. Starting from the SCM learned as above, we modified several assignment functions by replacing one argument by constant 1 (the middle category) and hence removing the according edge in the causal graph. We then used this SCM to resample data and to compute cc-Shapley values. We considered 7 SCMs arising from deleting the edges PKA $\to$ P38 , PKA $\to$ Jnk ,PKA $\to$ Mek, Mek $\to$ Erk, PKC $\to$ P38, PKC $\to$ Jnk and PKC $\to$ Raf respectively (Figures 15b-15h) and plotted the cc-Shapley values for the same features as in Figure 6b which we redraw in Figure 15a for comparison. We observe that deleting a causal edge between the target and a feature has, rather unsurprisingly, a strong impact on the relevance on the cc-Shapley values of the feature (cf. Figures 15b and 15c). Deleting an edge that is connected to a feature but not to the target has, in some cases, an impact on the cc-Shapley values of that feature (cf. Figure 15g). For the considered example, edges and features a removal of edges that are not connected to a feature did not lead to substantial changes in the cc-Shapley values for that feature (cf., e.g., Figure 15e).

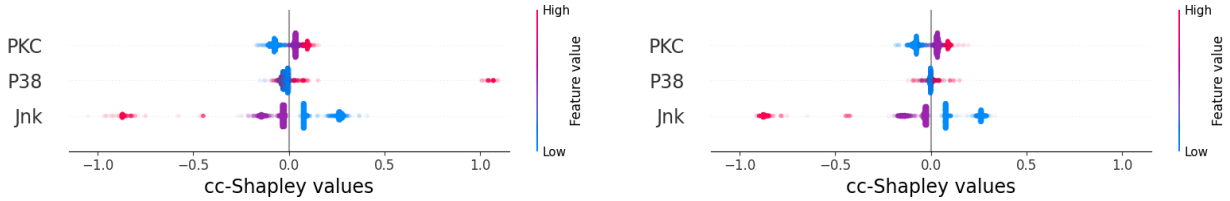


(a) univariate feature importances $I_\emptyset$



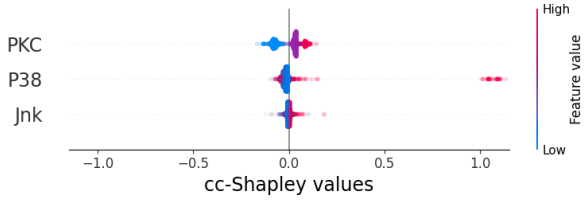
(b) Shapley values and cc-Shapley values

Figure 14: *Left*: univariate importance of all proteins considered in the experiment with the data from Sachs et al. [2005]. Standard errors (computed from 5 repetitions) are shown by shaded areas but are very small. *Right*: Shapley values and cc-Shapley values for all considered proteins.

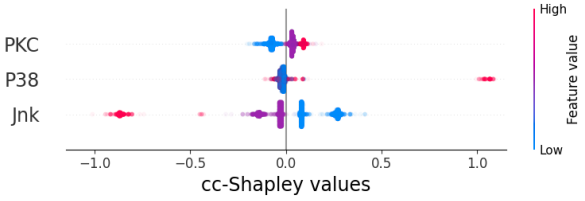


(a) original SCM (for comparison)

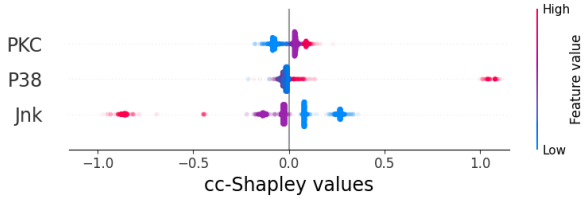
(b) without PKA → P38



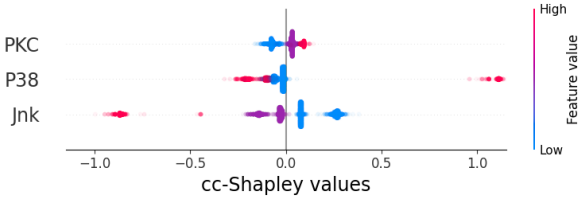
(c) without PKA → Jnk



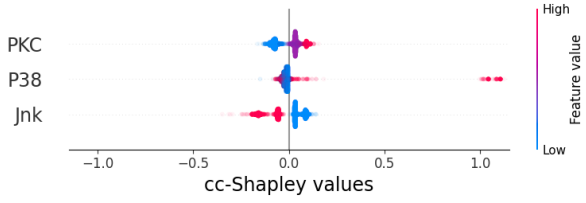
(d) without PKA → Mek



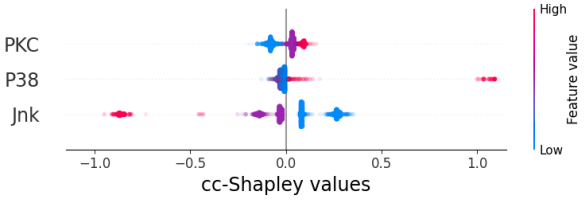
(e) without Mek → Erk



(f) without PKC → P38



(g) without PKC → Jnk



(h) without PKC → Raf

Figure 15: Sensitivity of Shapley and cc-Shapley values to the data-generating SCM. Figure 15a is identical to the right hand side of Figure 6b. Figures 15b-15h show the results of Shapley and cc-Shapley values for data generated from a modified SCM that arises from deleting the edge specified in the caption by setting the according argument in the assignment function to 1, cf. Section E.4 for details.