On Measuring Intrinsic Causal Attributions in Deep Neural Networks

Saptarshi Saha* SAPTARSHI.SAHA_R@ISICAL.AC.IN
Dhruv Vansraj Rathore CS2306@ISICAL.AC.IN
Soumadeep Saha SOUMADEEP.SAHA_R@ISICAL.AC.IN
Utpal Garain UTPAL@ISICAL.AC.IN
Indian Statistical Institute, Kolkata, West Bengal - 700108, India

David Doermann DOERMANN@BUFFALO.EDU
University at Buffalo, Buffalo, NY, USA

Editors: Biwei Huang and Mathias Drton

Abstract

Quantifying the causal influence of input features within neural networks has become a topic of increasing interest. Existing approaches typically assess direct, indirect, and total causal effects. This work treats NNs as structural causal models (SCMs) and extends our focus to include intrinsic causal contributions (ICC). We propose an identifiable generative post-hoc framework for quantifying ICC. We also draw a relationship between ICC and Sobol' indices. Our experiments on synthetic and real-world datasets demonstrate that ICC generates more intuitive and reliable explanations compared to existing global explanation techniques.

Keywords: Intrinsic Causal Contribution, Causal Normalizing Flow, Sobol Indices.

1. Introduction

In recent years, there has been a significant surge of interest in incorporating causal principles into deep learning models (Pawlowski et al., 2020; Saha and Garain, 2022). Much of the existing research has focused on post-hoc explanations of trained neural networks' decisions using causal effect analysis (Chattopadhyay et al., 2019; Alvarez-Melis and Jaakkola, 2017). Other studies have explored counterfactuals for explanations or data augmentation (Dash et al., 2022; Goyal et al., 2019b; Reddy et al., 2023b; Pitis et al., 2020), causal disentangled representation learning (Yang et al., 2021; Schölkopf et al., 2021; Shen et al., 2022), and causal discovery methods (Zhu et al., 2020). However, despite efforts (Chattopadhyay et al., 2019; Reddy et al., 2023a; Kancheti et al., 2021) to quantify the causal attributions learned by neural networks, there is presently no viable method for elucidating the "intrinsic causal contribution" (ICC) (Janzing et al., 2024) in neural networks. In this paper, we present a new framework based on generative models—the first of its kind, to the best of our knowledge—that quantifies intrinsic causal contributions in neural network models. To illustrate this concept of ICC, imagine a relay race with three runners: A, B, and C. Runner A starts the race and passes the baton late to runner B, who then hands it off late to runner C, who ends up finishing late as well. To determine the "intrinsic contribution" of runner B to the delay of runner C, we compare the delay of C to a situation where B only contributes the delay it inherited from A without adding any additional delay of its own. This means we're looking at how much delay B is responsible for beyond what it received from A. This concept helps differentiate between delays that B causes itself (intrinsic) and delays it simply passes on from A (inherited).

* Saptarshi was a Fulbright-Nehru Doctoral Research Fellow at the University at Buffalo during this work.

This distinction is meaningful whether we analyze the delay in a single race, the average delay across many races, or the variation in delays across multiple races.

To motivate the need for studying intrinsic causal contributions in neural network models, let's consider the task of predicting a patient's recovery time (R) using the features: treatment type (T), initial health condition (H), and post-treatment care (P). In the real world, H influences both T and P; while T also influences P. H, T, and P all influence R. However, these relationships among the input features H, T, and P are often not explicitly modeled in a neural network model. Now, assume that patients with severe initial health conditions are assigned to more aggressive treatment. It is possible that a neural network model might misattribute the longer recovery times directly to aggressive treatments without considering the severity of the initial health condition. Usual causal effect estimates the expected change in Recovery Time R as the treatment T changes. It doesn't account for the effect of upstream variable H on T (due to the do-intervention on T). With intrinsic attribution analysis, the model aims to understand the part of T's impact on R that is inherited from H, and the part that represents T's intrinsic effect. Thus, learning intrinsic causal attributions can also find application in medicine. For example, medical practitioners can look at treatments that have shown intrinsic benefits and consider optimizing these treatments for broader patient use.

To this end, the aim of our work is to identify the intrinsic causal contribution of an input on the output of a neural network. Our main contributions can be summarized as follows: We introduce an identifiable framework for computing intrinsic causal attributions in neural networks, a concept previously unexplored in neural network attribution to our knowledge. In addition to Shapley-based contributions, we advocate for asymmetric ICC. In Section 5, we demonstrate that ICC meets several desirable properties for an attribution method. In Section 6, we establish connections between the ICC and Sobol indices, offering a fresh perspective on global sensitivity analysis from a causal viewpoint. Finally, our experiments show that the ICC produces reliable global explanations.

2. Related works

Explainability Several established methods for explaining neural network models quantify the influence of input features on model outputs. These methods include saliency maps (Simonyan et al., 2014; Zeiler and Fergus, 2014; Selvaraju et al., 2017), Locally Interpretable Model-Agnostic Explanations (LIME) (Ribeiro et al., 2016a), Integrated Gradients (Sundararajan et al., 2017), DeepLift (Shrikumar et al., 2017), Shapley values (Lundberg and Lee, 2017) among others. While some of these techniques are model-agnostic, they are local in nature, meaning that the explanations are limited to individual predictions. On the other hand, global attributions are a powerful tool for interpretability because they highlight the importance of features across an entire population. They often use interpretable surrogate models like decision trees or adjust the input space to assess overall predictive power (Lakkaraju et al., 2016; Frosst and Hinton, 2017; Yang et al., 2018). Submodule pick LIME (SP-LIME) (Ribeiro et al., 2016b) uses submodular optimization to summarize local attributions, better capturing learned interactions. However, like surrogate models, it extracts useful and independent explanations from the LIME method, which may not effectively capture the non-linear feature interactions learned by neural networks. Ibrahim et al. (2019) proposed Global Attribution Mapping (GAM) to explain the non-linear representations learned by a neural network across different subpopulations. GAM clusters similar local feature importances to create human-interpretable global attributions, each tailored to explain a specific subpopulation. Additionally, GAM allows for adjustable granularity to capture varying numbers of subpopulations in its global

explanations. Permutation Feature Importance (PFI) (Breiman, 2001; Strobl et al., 2008) is another comparable measure across model types, offering a global view of the model's reliance on each feature. However, none of these methods account for causality in their explanations.

Causal Explanations Frye et al. (2021) proposed Asymmetric Shapley Values to integrate real-world causal knowledge by restricting feature permutations to those that align with a (partial) causal ordering. Heskes et al. (2020) introduced causal Shapley values that account for the causal relationships among features to explain their total causal effect on predictions. Jung et al. (2022) presented the do-Shapley values to measure the strengths of different causes to a target quantity. Chattopadhyay et al. (2019) proposed a post-hoc explanation method to find average causal effects in a trained neural network by treating it as an structural causal model (SCM). It prompts further studies (Yadu et al., 2021; Wang et al., 2023; Schwab and Karlen, 2019; Goyal et al., 2019a) to quantify learned causal effects more comprehensively. Reddy et al. (2023a) introduce an ante-hoc method that identifies and retains direct, indirect, and total causal effects during the neural network model training process. Other causal explanation methods (Verma et al., 2022; Goyal et al., 2019b; Wachter et al., 2018; Dandl et al., 2020; Kommiya Mothilal et al., 2021; Mahajan et al., 2019; Van Looveren and Klaise, 2021) leverage counterfactuals to examine model behavior under semantically meaningful input changes. Breuer et al. (2024) propose a causality-aware, model-agnostic framework based on Shapley values for global explanations. However, none of the existing work attempts to quantify ICC for attributions in deep neural networks.

Sensitivity Analysis Sensitivity analysis (SA) (Saltelli et al., 2008) studies how model inputs influence outputs and is widely used to explain input-output relationships in complex systems. Scholbeck et al. (2024) argue that interpretable machine learning is essentially a form of sensitivity analysis applied to machine learning models. FEL et al. (2021) used Sobol' indices to model the attributions of image regions. Kuhnt and Kalka (2022); Stein et al. (2022) and Scholbeck et al. (2024) present an overview of sensitivity analysis methods for interpreting ML models. Tunkiel et al. (2020) apply derivative-based sensitivity analysis to rank high-dimensional features in a directional drilling model. Stein et al. (2022) use the Morris method to calculate sensitivity indices for genomic prediction. Bénesse et al. (2024) demonstrate how fairness can be defined within a global sensitivity analysis (GSA) framework, highlighting shared indicators between the two fields. They also demonstrate how GSA frameworks can address causal fairness, using specific Sobol' indices to detect causal links between sensitive variables and algorithm outcomes. The generalization of Sobol indices within a causal framework remains largely unexplored.

3. NNs through the lens of SCMs

Notation Each random variable is denoted by an uppercase letter (e.g., V) and its realized value by the corresponding lowercase letter (e.g., v). We use boldface letters $\mathbf{V}$ and $\mathbf{v}$ to represent a set of variables and their corresponding realized values, respectively. The set $\{1, \ldots, p\}$ is denoted as $[p]$. As we often need to work with $A \cup \{j\}$, it is handy to write $A + j$ for it. $A - j$ represents the set difference $A \setminus \{j\}$. Throughout this work, we use P to denote probability distributions and $\tilde{p}$ to represent the corresponding density or probability mass functions (e.g., $P(X_1)$ vs. $\tilde{p}(x_1)$).

This work is grounded in the principles of causality, specifically SCMs and the do-calculus, as outlined by Pearl (2009). A concise overview of the relevant concepts is provided in the Appendix A. Consider a causal graph $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, where $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ represents the set of input

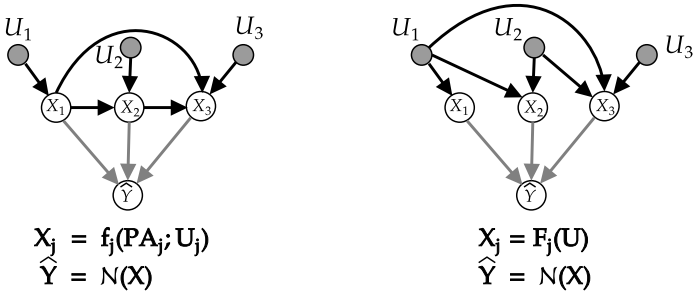

Figure 1: An example of a causal view of a NN with three input features. White nodes represent variables that are either observed or assumed to be known, while shaded nodes indicate unobserved or latent variables. The left graph illustrates the causal relationships between features along with their exogenous parents, while the right graph utilizes exogenous variables for the TMI mapping of the SCM of inputs. In both figures, the grey edges serve to augment the neural network to the SCM.

features (random variables), and $\mathcal{E}$ denotes the set of edges that capture the causal relationships among the variables in $\mathbf{X}$.

Assumption 1 *The causal graph $\mathcal{G}$ is acyclic and contains no latent (unobserved) confounders.*

Let $\mathcal{N}$ be a neural network model that has been trained to predict Y from input features $\mathbf{X}$ by minimizing the empirical loss. The neural network $\mathcal{N}$ can be envisioned as a directed acyclic graph (DAG) consisting of directed edges that link successive layers of neurons. Consequently, the predicted output $\hat{Y} = \mathcal{N}(\mathbf{X})$ can be interpreted as the outcome of a sequence of interactions from the initial layer to the final layer of the network $\mathcal{N}$. When analyzing the intrinsic contributions of inputs on the output of $\mathcal{N}$, only the neurons in the first and final layers are considered. Therefore, akin to the approach in Chattopadhyay et al. (2019); Kancheti et al. (2021) we can marginalize the influence of the hidden layers within $\mathcal{N}$ and concentrate exclusively on the causal structure between inputs and outputs. With our view of a neural network as an SCM, we define augmented causal graph $\tilde{\mathcal{G}} = (\mathcal{V}, \tilde{\mathcal{E}})$ with $\mathcal{V} = \mathbf{X} \cup \{\hat{Y}\}$ and $\tilde{\mathcal{E}} = \mathcal{E} \cup \bigcup_{j=1}^{p} \{(X_j, \hat{Y})\}$. Note that while our perspective on neural networks as SCMs is the same as Chattopadhyay et al. (2019); Kancheti et al. (2021), they do not address or model intrinsic causal attribution, which is central to our study. To measure the intrinsic contribution of each feature to $\hat{Y}$, we first recursively substitute structural equations into one another, expressing each feature X_j solely in terms of the unobserved noise variables $\mathbf{U}$:

$$X_j = f_j(PA_j; U_j) = F_j(\mathbf{U}) = F_j(U_1, ..., U_p), \qquad \forall 1 \leq j \leq p. \tag{1}$$

Figure 1 portrays an example of our SCM perspective on neural networks. As $\mathcal{G}$ is acyclic, $\mathbf{F} = (F_1, F_2..., F_p)$ is a triangular map. More importantly, any SCM can be represented as a tuple $(\mathbf{F}, P_{\mathbf{U}}) \in \mathcal{F} \times \mathcal{P}_{\mathcal{U}}$, where $\mathcal{F}$ denotes the set of all triangular monotonic increasing (TMI) maps, and $\mathcal{P}_{\mathcal{U}}$ represents the set of all fully-factorized distributions $P_{\mathbf{U}}(\mathbf{u}) = \prod_{j=1}^{p} P_{U_j}(u_j)$. TMI maps are autoregressive functions where the i-th component is strictly monotonically increasing with respect to its i-th input. Mathematically, a TMI map is characterized as a function $T : \mathbb{R}^p \to \mathbb{R}^p$ defined as

follows:

$$T(u) = \begin{bmatrix} T_1(u_1) & T_2(u_1, u_2) & \cdots & T_d(u_1, \ldots, u_p) \end{bmatrix}^\top,$$

where each component function T_k depends solely on the first k variables $u_{\leq k} := (u_1, \ldots, u_k)$ and is monotone increasing with respect to the last input u_k for any $(u_{k+1}, \ldots, u_p)$. They can approximate any fully supported distribution and can be parameterized through deep neural networks. For more details on TMI maps, please see Xi and Bloem-Reddy (2023), Irons et al. (2021).

4. Intrinsic Causal Contributions

Now that we can conceptualize each X_j as being influenced by the independent causal factors $U_1, ..., U_p$, the change in uncertainty within $\hat{Y}$ is assessed as a result of a hypothetical adjustment of U_j (which is standard conditioning stemming from exogeneity). We initiate our discussion by defining the intrinsic causal contributions (Janzing et al., 2024) of input features to the output of a trained neural network model $\mathcal{N}$.

Definition 1 (Intrinsic Causal Contribution in Neural Network)
Given an adjustment set $I \subseteq [p] - \{j\}$, the intrinsic causal contribution of a feature X_j on the output $\hat{Y}$ of a NN $\mathcal{N}$ is defined as

$$ICC_\phi(X_j \to \hat{Y}|I) = \phi(\hat{Y}|\mathbf{U}_{j+I}) - \phi(\hat{Y}|\mathbf{U}_I), \tag{2}$$

where ϕ is a measure of conditional uncertainty that satisfies one of the following two conditions:

1. ***monotonicity:*** $\phi(\hat{Y}|\mathbf{U}_I) \geq \phi(\hat{Y}|\mathbf{U}_{j+I})$, *or* $\phi(\hat{Y}|\mathbf{U}_I) \leq \phi(\hat{Y}|\mathbf{U}_{j+I})$
2. ***calibration:*** $\phi(\hat{Y}) = \phi(\hat{Y}|\mathbf{U}_\varnothing) = 0$, *or* $\phi(\hat{Y}|\mathbf{U}) = 0$

where $\varnothing$ is the empty set. In this context, $\phi(\cdot|U_I)$ denotes conditioning on all noise variables U_i for which i is an element of the set I.

It is important to note that $\hat{Y}$ is a deterministic function of $\mathbf{X}$, which differs very slightly from Janzing et al. (2024). We adapt the definition in the context of explaining the decisions made by a neural network. Monotonicity is not an absolute necessity, but it is often simpler to understand and work with positive contributions in practical scenarios. For example, the variance of the conditional expectation can be seen as a choice of ϕ. We will discuss this in more detail later. The definition of ICC can be generalized to scenarios involving confounding variables. However, we will not discuss this here to maintain focus on the main discussion. Please look at the foundational ICC paper Janzing et al. (2024) for more details.

4.1. Measuring ICC using Shapley value and topological ordering

Unfortunately, the contribution of each feature X_j in (2) depends on the adjustment set I given as context. Janzing et al. (2024) address this issue by using Shapley values to symmetrize the ICC:

$$ICC_\phi^{\text{Sh}}(X_j \to \hat{Y}) = \sum_{T \subseteq [p]-j} \frac{1}{p\binom{p-1}{|T|}} ICC_\phi(X_j \to \hat{Y}|T). \tag{3}$$

However, computing the Shapley values can be computationally expensive. Another viable option would be to utilize topological ordering. A topological ordering of a DAG $\mathcal{G}$ is a specific arrangement of the its d nodes such that each node is positioned earlier in the sequence than any of its descendants. Let S_p be the symmetric group on the set $[p]$. Given an ordering $\pi \in S_p$, let T_π^j be the set of indices that occur before j in the ordering π, i.e., $T_\pi^j = \{k : \pi(k) < \pi(j)\}$. If we define

$$ICC_\phi^\pi(X_j \to \hat{Y}) = ICC_\phi(X_j \to \hat{Y}|T_\pi^j), \tag{4}$$

it is easy to see that the reliance on arbitrary π introduces an unnecessary level of ambiguity. Therefore, we may constrain π as a topogical (or causal) ordering of the DAG $\mathcal{G}$. However, several valid sequences can meet this requirement, making the ordering non-unique. And as π is not unique, the ambiguity remains unresolved. Therefore, we adjust (4) by averaging exclusively over all potential topological orderings of the DAG $\mathcal{G}$:

$$ICC_\phi^{\text{To}}(X_j \to \hat{Y}) = \frac{1}{|\mathcal{C}(\mathcal{G})|} \sum_{\pi \in \mathcal{C}(\mathcal{G})} ICC_\phi(X_j \to \hat{Y}|T_\pi^j), \tag{5}$$

where $\mathcal{C}(\mathcal{G})$ is set of all possible causal ordering of $\mathcal{G}$. An iterative adaptation of the algorithm proposed by Knuth and Szwarcfiter (1974) can be used to generate all possible topological orderings. The Shapley-based ICC can also be expressed in an alternative, equivalent form (Mitchell et al., 2022):

$$ICC_\phi^{\text{Sh}}(X_j \to \hat{Y}) = \frac{1}{p!} \sum_{\pi \in S_p} ICC_\phi(X_j \to \hat{Y}|T_\pi^j). \tag{6}$$

As $|\mathcal{C}(\mathcal{G})| \leq |S_p|$, it is clear that from equations 5 and 6 that ICC_ϕ^{To} is computationally more efficient than ICC_ϕ^{Sh}. We define the value of the coalition for any $T \subseteq [p]$ as $\phi(\hat{Y}|do(\mathbf{X}_T)) := \sum_{\mathbf{x}_{T_\pi^j}} \phi\big(\hat{Y}|do(\mathbf{X}_T = \mathbf{x}_T)\big)\tilde{p}(\mathbf{x}_T)$.

Lemma 1 (Janzing et al. (2024)) *For a topological ordering $\pi \in \mathcal{C}(\mathcal{G})$, we have*

$$\phi(\hat{Y}|\mathbf{U}_{T_\pi^j}) = \phi(\hat{Y}|\mathbf{X}_{T_\pi^j}) = \phi(\hat{Y}|do(\mathbf{X}_{T_\pi^j})).$$

Instead of criticizing ICC_ϕ^{To} for avoiding rung 3 causal models (Janzing et al., 2024, Section 5), we recognize that Lemma 1 is crucial for establishing the identifiability of the ICC_ϕ^{To}. This means we can measure the causal contribution (ICC_ϕ^{To}) solely using observational conditionals that are readily estimable from observational data. We do not see ICC_ϕ^{To} as a replacement for ICC_ϕ^{Sh}. The intrinsic value of the latter is undeniable. Instead, we view both formulations as suited for different contexts. Due to space constraints, we provide the causal interpretation of ICC in the Appendix B. We also recommend that readers refer to Janzing et al. (2024) for further insights. Hereafter, when we refer to ICC, it encompasses both ICC_ϕ^{To} and ICC_ϕ^{Sh}.

5. Mathematical properties of ICC

This Section demonstrates that intrinsic causal contributions satisfy several desirable properties (Janzing et al., 2020). To the best of our knowledge, our effort here is the first attempt at an axiomatic characterization of ICC.[1] Due to space limitations, all proofs are provided in the Appendix D.

1. Jung et al. (2022) highlight the complete characterization of ICC as an open problem.

Property 1 (Efficiency/ Completeness)

$$\sum_j ICC_\phi^{To}(X_j \to \hat{Y}) = \phi(\hat{Y}|\mathbf{U}) - \phi(\hat{Y}).$$

In the context of neural network attribution, efficiency means that the uncertainty in the model's output is fully distributed across its input features. Shapley-based ICC values inherently guarantee the completeness, owing to the general properties of Shapley values (Shapley, 1953).

Property 2 (Nullity/ Dummy)

$$ICC_\phi^{To}(X_j \to \hat{Y}) = ICC_\phi^{Sh}(X_j \to \hat{Y}) = 0$$

whenever $\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}})$ *for all* $I \subseteq [p] - \{j\}$.

Nullity ensures that if a feature is entirely disconnected from the model's output, it receives no contribution.

Property 3 (Symmetry)

$$ICC_\phi^{Sh}(X_j \to \hat{Y}) = ICC_\phi^{Sh}(X_l \to \hat{Y}),$$

if $\phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}}) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{l\}}) \quad \forall I \ \ s.t. \ \ I \subseteq [p] - \{j, l\}$.

Symmetry requires that attribution be equally distributed among features that provide the same information for the model's prediction. While ICC_ϕ^{Sh} satisfies the symmetry property, ICC_ϕ^{To} does not. However, the symmetry property is not without controversy, as symmetrical approaches to model explainability can obscure known causal relationships in the data (Frye et al., 2021).

Property 4 (Sensitivity/ Causal irrelevance) *If* X_i *is causally irrelevant (Galles and Pearl, 1997) to* $\hat{Y}$ *for all* $I \subseteq [p] - \{i\}$*, i.e.,*

$$P(\hat{Y}|do(X_i, \mathbf{X}_I)) = P(\hat{Y}|do(\mathbf{X}_I)), \quad \forall I \subseteq [p] - \{i\},$$

then

$$ICC_\phi^{Sh}(X_i \to \hat{Y}) = ICC_\phi^{To}(X_i \to \hat{Y}) = 0.$$

Causal irrelevance captures the causes of an outcome by ensuring that variables not related to the outcome have zero contribution. From a causal viewpoint, it is also related to sensitivity (Sundararajan et al., 2017): if the function implemented by the deep network does not mathematically depend on a particular variable, then the attribution for that variable should always be zero. Implementation invariance (Sundararajan et al., 2017) axiom loses significance if it refers to the properties of functions rather than focusing on the properties of algorithms (Janzing et al., 2020). While linearity is often a desirable property in many attribution methods, recent progress has been towards non-linear attribution methods (FEL et al., 2021). The linearity of ICC depends on the choice of the function ϕ. However, we sacrifice linearity by focusing on using a variance-based uncertainty measure as a candidate for ϕ.

Choice of ϕ. So far, we have considered ϕ as a general measure of uncertainty without specifying its form. However, we need to adopt a suitable ϕ for practical purposes. While Janzing et al. (2024) suggest using variance and entropy for contribution analysis, we will focus on variance-based uncertainty measures in this article. Quantifying uncertainty using variance is often more intuitive and easier to estimate from finite data. Furthermore, variance-based measures meet several desirable properties (axioms) for assessing second-order uncertainty (Corbière et al., 2021), as discussed by Sale et al. (2023). They have been proposed as an alternative to entropy-based measures, which have recently faced criticism in the literature(Wimmer et al., 2023). By defining $\phi(\hat{Y}|\mathbf{U}_I) := \mathbb{V}_{\mathbf{U}_I}(\mathbb{E}(\hat{Y}|\mathbf{U}_I))$, we can express the contribution of variable X_j to $\hat{Y}$, given the context I, as:

$$ICC_\phi(X_j \to \hat{Y}|I) = \mathbb{V}_{\mathbf{U}_{I+j}}(\mathbb{E}(\hat{Y}|\mathbf{U}_{I+j})) - \mathbb{V}_{\mathbf{U}_I}(\mathbb{E}(\hat{Y}|\mathbf{U}_I))$$

The difference between two variances allows us to measure the intrinsic contribution of X_j to the uncertainty in predicting $\hat{Y}$, relative to the context I. The monotonicity of the variance of the conditional expectation is immediate from the following theorem.

Theorem 2 *Let X, Y and Z be random variables on the same probability space and $\mathbb{V}(X) < \infty$. Then,*

$$\mathbb{V}_Y(\mathbb{E}(X|Y)) \leq \mathbb{V}_{Y,Z}(\mathbb{E}(X|Y,Z)).$$

Note that while Theorem 2 is stated for random variables, it also holds for random vectors $\mathbf{Y}, \mathbf{Z}$. We have hitherto guaranteed that contributions are positive. However, normalized contributions are easier to interpret and visualize. The decomposition of the total variance of $\hat{Y}$ provides a natural way for normalizing the ICC:

Corollary 3 (Causal Decomposition of Variance) *Let $\mathbb{V}(\hat{Y}) < \infty$. Then,*

$$\mathbb{V}(\hat{Y}) = \sum_{j=1}^{p} ICC_\phi^{To}(X_j \to \hat{Y}) = \frac{1}{|\mathcal{C}(\mathcal{G})|} \sum_{j=1}^{p} \sum_{\pi \in \mathcal{C}(\mathcal{G})} \Big(\mathbb{V}\big(\mathbb{E}(\hat{Y}|\mathbf{X}_{T_\pi^j + j})\big) - \mathbb{V}\big(\mathbb{E}(\hat{Y}|\mathbf{X}_{T_\pi^j})\big) \Big).$$

The proof of the Corollary 3 is immediate from Property 1 and our choice of ϕ. A similar variance decomposition is straightforward for Shapley-based ICC. Going forward, we will consider ϕ to be normalized by $\mathbb{V}(\hat{Y})$. i.e., $\phi(\hat{Y}|\mathbf{U}_I) := \frac{\mathbb{V}_{\mathbf{U}_I}(\mathbb{E}(\hat{Y}|\mathbf{U}_I))}{\mathbb{V}(\hat{Y})}$.

6. Comparison with Sobol' method

While variance-based sensitivity analysis (Sobol', 2001) accommodates non-linear models, it falls short of capturing causal influence. This is because it focuses on reducing variance by conditioning on observed variables without distinguishing whether the statistical relationship with the target is causal or merely confounded. We view corollary 3 as a causal decomposition of variance as it allows for the partial allocation of the output variance to each input variable while respecting the causal ordering, thereby generalizing classical functional ANOVA decomposition of variance within a causal framework. Although efforts (Li et al., 2010; Kucherenko et al., 2012; Rahman, 2014; Hooker, 2007) have been made to generalize ANOVA by removing the independence assumption among input variables, none have addressed this issue from a causal perspective. Finally, we establish the connection between variance-based ICC and Sobol' indices, assuming independent input variables, in the following theorem:

Theorem 4 *Assume that input features $\{X_i\}$ are independent. Then, with our specific choice of ϕ, for any $I \subseteq [p]$ and any $j \in [p]$, we have*

$$\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{X}_I) = \sum_{T \subseteq I} \mathcal{S}_T; \qquad ICC_\phi(X_j \to \hat{Y}) = \sum_{j \in T, T \subseteq [p]} \frac{\mathcal{S}_T}{|T|}, \tag{7}$$

where $\mathcal{S}_T$ is the Sobol' sensitivity index of the input subset $\mathbf{X}_T$ (see Appendix C for more details).

Theorem 4 reflects that ICC may be viewed as a step toward generalizing Sobol indices within a causal framework.

7. Learning and explaining ICC in NNs

In this Section, we introduce our primary post-hoc (Retzlaff et al., 2024) approach to identifying and explaining the intrinsic causal contributions of an input feature. We will start by formally defining the concept of identifiability within the context of ICC.

Definition 5 *For a given neural network $\mathcal{N}$, the intrinsic causal contribution of an input feature X_j on the output $\hat{Y}$ is identifiable if $ICC_\phi(X_j \to \hat{Y})$ can be computed uniquely from any positive probability distribution $P(\mathbf{X}, \hat{Y})$.*

Prior work (Janzing et al., 2024) does not address the issue of identifiability for ICC. Under the assumption of no latent confounding (Assumption 1), and based on Lemma 1, it is straightforward that ICC_ϕ^{To} is identifiable. However, in general, we are unable to find a way to compute $\phi(\hat{Y}|\mathbf{U}_I)$ without knowledge of the SCM. For example, Janzing et al. (2024) inferred the SCM based on common sense knowledge and assigned all regression coefficients a value of 1. For a dataset (Quinlan, 1993) with non-linearities, they applied an additive noise model for a simple approximation of structural equations. In contrast, we propose to learn the entire causal-generating process using causal normalizing flow (CNF) (Javaloy et al., 2023) as they are a natural choice for approximating a wide range of causal data-generating processes. Nevertheless, generative models are vulnerable to cases where the latent values ($\mathbf{u}$) underlying observations cannot be determined uniquely (Khemakhem et al., 2019), no matter how much empirical data is available, which may lead to inaccurate estimation of $\phi(\hat{Y}|\mathbf{U}_I)$. In this scenario, we guarantee that, even though different but equivalent model (CNF) fits may be obtained from the same data, the estimation of ϕ remains consistent, provided the following assumptions hold. For further details on normalizing flows, see Papamakarios et al. (2021).

Assumption 2 *We constrain the class of SCMs under consideration by adopting the following fairly common assumptions from Javaloy et al. (2023): i) the data-generating process is diffeomorphic — that is, $\mathbf{F}$ is invertible, and both $\mathbf{F}$ and its inverse are differentiable; ii) causal sufficiency, i.e., $P_{\mathbf{U}}(\mathbf{u}) = \prod_{j=1}^{p} P_{U_j}(u_j)$.*

Causal normalizing flows are themselves parametric TMI maps that can approximate any other TMI map with arbitrary precision. With SCMs and causal NFs categorized under the same family, we leverage existing results on identifiability (Xi and Bloem-Reddy, 2023).

Theorem 6 (Javaloy et al. (2023), Xi and Bloem-Reddy (2023)) *If two elements from the family $\mathcal{F} \times \mathcal{P}_{\mathcal{U}}$ yield the same observational distribution, then their data-generating processes differ only by a component-wise (Borel measurable) invertible transformation of the variables $\mathbf{U}$.*

Theorem 6 says that if a causal normalizing flow $(\mathbf{F}_\theta, P_\theta) \in \mathcal{F} \times \mathcal{P}_\mathcal{U}$ matches the observational distribution generated by $(\mathbf{F}, P_\mathbf{U}) \in \mathcal{F} \times \mathcal{P}_\mathcal{U}$, then the exogenous variables in the flow differ from the true exogenous variables only through independent, component-wise invertible transformations. Mathematically, for $\mathbf{U} \sim P_\mathbf{U}$, it holds that $\mathbf{F}_\theta^{-1}(\mathbf{F}(\mathbf{U})) \sim P_\theta$ and $\mathbf{F}_\theta^{-1}(\mathbf{F}(\mathbf{u})) = \mathbf{h}(\mathbf{u}) = (h_1(u_1), h_2(u_2), ..., h_p(u_p))$, where each h_i is an invertible function. This component-wise invertibility is fundamental to the identifiability of ICC:

Theorem 7 *Under Assumption 2, suppose we have two CNFs $(\mathbf{F}_{\theta_1}, P_{\theta_1})$ and $(\mathbf{F}_{\theta_2}, P_{\theta_2})$ that both match the observational distribution $P(\mathbf{X}, \hat{Y})$, then the intrinsic causal contributions of X_j on $\hat{Y}$ will be equal for both CNFs. Specifically, we have:*

$$ICC_{\phi,\theta_1}(X_j \to \hat{Y}) = ICC_{\phi,\theta_2}(X_j \to \hat{Y}).$$

Another crucial advantage of using CNF framework (Javaloy et al., 2023) is its ability to handle both mixed continuous-discrete data and partial knowledge of the causal graph, making it highly applicable to real-world scenarios. To handle discrete data, we adopt the general method by Xi and Bloem-Reddy (2023) that transforms the observed discrete variables into continuous ones by adding independent noise $\epsilon \in [0, 1]$ —such as standard uniform noise — ensuring the original distribution remains recoverable. Essentially, we posit that discrete variables represent the integer parts of noisy continuous variables generated under an SCM that meets our assumptions. Thereby, it allows our theoretical and practical insights to remain applicable. Recently, Si et al. (2023) have raised questions about using likelihood loss to train normalizing flows. In line with their approach, we use MMD loss instead of likelihood in our experiment to train the normalizing flows. In training the flow model, we focus on the crucial step of estimating ϕ, which is essential for computing the intrinsic causal contributions.The computation of $\mathbb{V}_{\mathbf{U}_I}(\mathbb{E}_{\mathbf{U}_{-I}}(\hat{Y}|\mathbf{U}_I))$ involves a two-fold integration, which could be challenging. We therefore present a Monte Carlo-based algorithm for the efficient estimation of ϕ.

Algorithm 1 Pseudocode for estimating ϕ

Input: Batch size B, context I for conditioning, trained CNF $(\mathbf{F}_\theta, P_\theta)$, the neural network $\mathcal{N}$

$\mathbf{u}_M^{(i)}, \mathbf{u}_N^{(i)} \sim P_\theta$ for $i = 1, 2, ..., B$.

$\mathbf{u}_Q^{(i)} = \left(\mathbf{u}_{M_{-I}}^{(i)}, \mathbf{u}_{N_I}^{(i)}\right)$ for $i = 1, 2, ..., B$.

$\hat{y}_M^{(i)} = \mathcal{N}\Big(\underbrace{\mathbf{F}_\theta(\mathbf{u}_M^{(i)})}_{\mathbf{x}_M}\Big), \hat{y}_N^{(i)} = \mathcal{N}\Big(\underbrace{\mathbf{F}_\theta(\epsilon_V)}_{\mathbf{x}_N}\Big), \hat{y}_Q^{(i)} = \mathcal{N}\Big(\underbrace{\mathbf{F}_\theta(\epsilon_W)}_{\mathbf{x}_Q}\Big)$ for $i = 1, 2, ..., B$.

$\bar{y} = \frac{1}{2B} \sum_{i=1}^{B} \left(\hat{y}_M^{(i)} + \hat{y}_N^{(i)}\right); V = \frac{1}{2B-1} \sum_{i=1}^{B} \left((\hat{y}_M^{(i)} - \bar{y})^2 + (\hat{y}_N^{(i)} - \bar{y})^2\right)$

$\hat{\psi} = V - \frac{1}{2B} \sum_{i=1}^{B} \left(\hat{y}_N^{(i)} - \hat{y}_Q^{(i)}\right)^2$

Output: $\hat{\phi} = \frac{\hat{\psi}}{V}$

In Algorithm 1, we employ the Jansen estimator (Jansen, 1999), widely recognized as one of the most efficient (Puy et al., 2022). Jansen's method is commonly employed alongside a Monte Carlo sampling strategy. We improve upon the standard Monte Carlo method by employing a Randomized Quasi-Monte Carlo (RQMC) sampling strategy (L'Ecuyer, 2018), which generates low-discrepancy

sample sequences for faster and more stable convergence rates (L'Ecuyer and Lemieux, 2002; Gerber, 2015). RQMC methods enable them to be considered variance reduction techniques for the standard Monte Carlo method. Scrambled nets, a type of RQMC, offer valuable robustness properties (Owen and Rudolf, 2021). Our experiments utilize the most commonly used QMC method: Sobol' sequences (Sobol', 1967), which can be scrambled (Owen, 1998). Although Algorithm 1 could be applied to both ICC^{To} and ICC^{Sh}, for the sake of completeness, we provide an algorithm in the Appendix E — using Lemma 1 — specifically dedicated to computing ϕ for ICC^{To}, where the CNF is not necessary.

8. Experiment and analysis

Now, we demonstrate that ICCs provide a natural framework for global explanations. We perform experiments on three datasets: a synthetic dataset and two well-known real-world benchmark datasets, AutoMPG (Quinlan, 1993) and COMPAS (Larson et al., 2016). The causal graph of these datasets is depicted in Figure 2. Appendix F shows more detailed information on each dataset. We compare the ICC with global attributions generated by GAM, SP-LIME, and permutation feature importance (PFI). We apply GAM to five different local attribution methods: Integrated Gradients (IG), Gradient $\times$ Inputs (I$\times$G), SmoothGrad (SG), Shapley Values (SHAP), and LIME — to generate global attributions for the test samples. We train a three-layer feed-forward neural network with ReLU activation functions on each of the three balanced datasets. The performance metrics of these networks are presented in Table 1. To calculate the ICC for each dataset, we fit a CNF to approximate the SCM of the input features. Each CNF is constructed using Masked Autoregressive Flows (Papamakarios et al., 2017) as its layers. We assess the quality of these flows using the 1-Wasserstein distance metric, with the results reported in Table 1. We compute attribution scores on the test dataset. We use the OpenXAI (Agarwal et al., 2022) codebase as the foundation for our implementation.

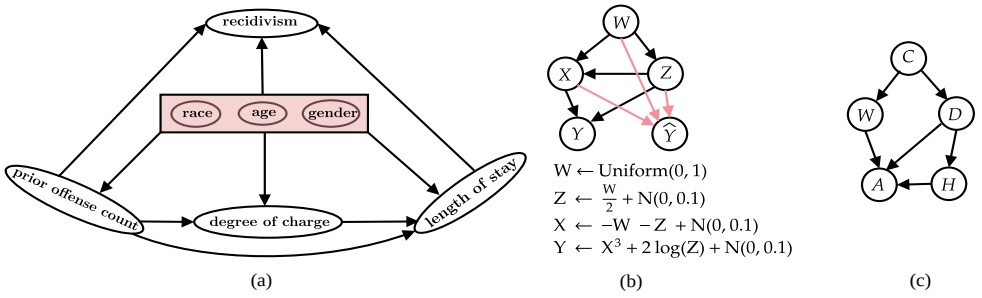

Figure 2: Causal graphs for experimental datasets: (a) COMPAS, (b) Synthetic, (c) AutoMPG

Global explanation research inherently struggles with effective and reliable validation due to the absence of baseline truths for attributions, making identifying appropriate validation methodologies an open research question. In the absence of a ground truth, we evaluate the reliability of an attribution method by adapting the Prediction Gap on Unimportant feature perturbation (PGU) (Dai et al., 2022; Petsiuk et al., 2018). This metric measures the change in the network's output when unimportant features are set to zero, while the (top-k) influential features identified by a post

Table 1: **Left:** Performance metrics for $\mathcal{N}$ to Be Explained (Root Mean Squared Error (RMSE) for Regression and F_1 score for Classification). **Right:** Quality metrics (1-Wasserstein Distance) of CNF models used to compute the ICC.

$\mathcal{N}$	**RMSE**(↓)	$\mathbf{F_1}$ **Score** (↑)
Synthetic	$0.1024_{\pm 0.0019}$	N/A
Auto-MPG	$0.1103_{\pm 0.0032}$	N/A
COMPAS	N/A	$0.9115_{\pm 0.0016}$

CNF Model	$\mathcal{W}_1$**-Distance**
Synthetic	$0.5553_{\pm 0.0028}$
Auto-MPG	$0.9641_{\pm 0.0114}$
COMPAS	$0.9562_{\pm 0.0398}$

hoc explanation remain unchanged. Smaller values on this metric indicate higher reliability in the explanation. For each dataset, the PGU values for every attribution method are reported in Table 2.

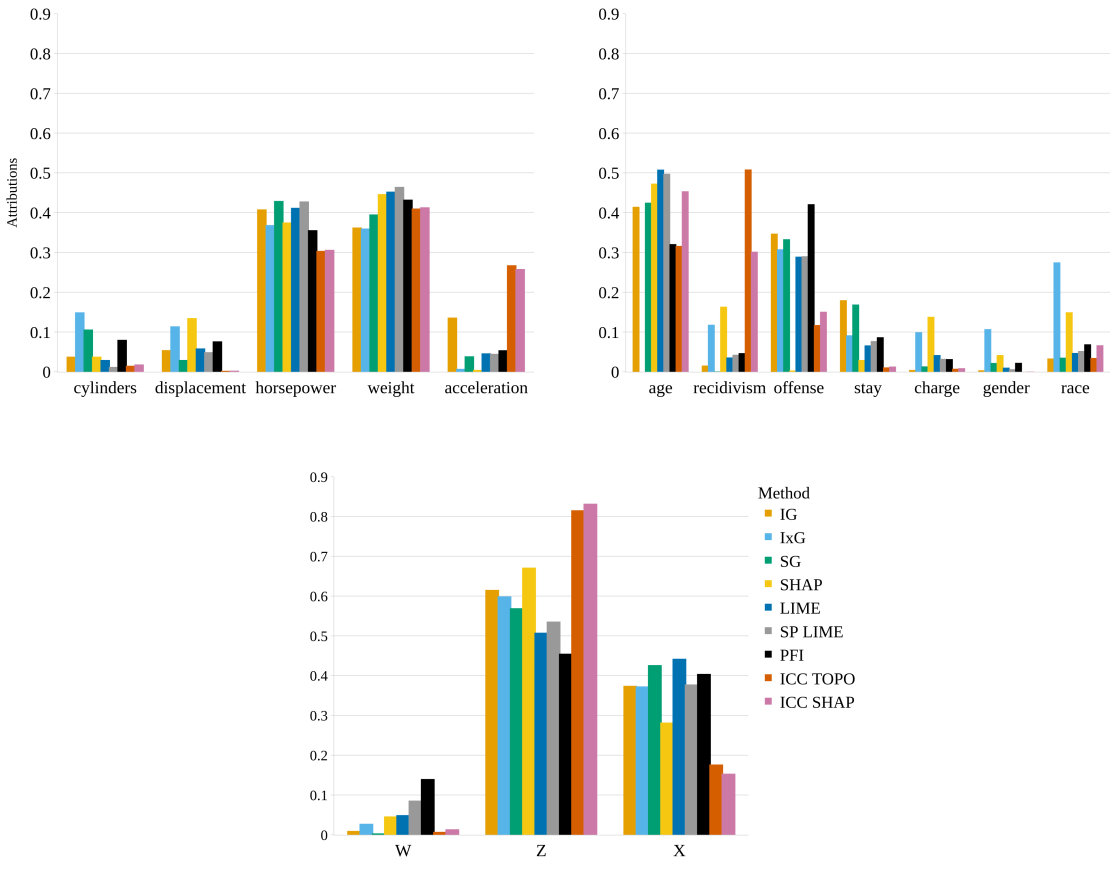

Figure 3: Global attribution explanations - feature importances. Top left: AutoMPG dataset. Top right: COMPAS dataset. Bottom: Synthetic dataset.

Figure 3 presents the attribution values of the input features. In the synthetic dataset, we observe that both ICC methods assign negligible attribution to W, which aligns with expectations as the true outcome Y depends on W only through X and Z. SmoothGrad and IG methods also support this assignment. Similarly, in the AutoMPG experiment, both ICC methods produce comparable attributions. However, a discrepancy appears in the COMPAS experiment between ICC^{To} and ICC^{Sh}. Specifically, ICC^{To} identifies recidivism as the most critical feature, while ICC^{Sh} ranks it as the second most important. More notably, most other methods assign relatively low importance (attribution ≤ 0.25) to recidivism, focusing instead on prior offense count as the more influential attribute. To further examine this discrepancy, we train separate classifiers using each feature individually: age, recidivism, and prior offense count. The resulting F_1 scores are 0.8972, 0.8964, and 0.8912, respectively.

Table 2: PGU(↓) values for different datasets (scaled by 1×10^{-1}). We report the aggregated PGU by summing across all values of k.

Dataset	IG	I×G	SG	SHAP	LIME	SP-LIME	PFI	ICC^{To}	ICC^{Sh}
Synthetic	1.9	1.9	1.9	1.9	1.9	4.77	1.9	1.9	1.9
Auto MPG	2.84	3	2.86	2.78	2.72	6.48	2.76	**2.52**	**2.52**
COMPAS	5.73	18.29	6.47	8.51	6.06	6.36	5.48	5.18	**5.17**

9. Conclusion

This paper proposes a framework that leverages Intrinsic Causal Contributions to generate global attributions that complement existing interpretability techniques for neural networks. Additionally, we establish a link between classical sensitivity analysis and Intrinsic Causal Contributions that bridge causality and sensitivity analysis. This connection suggests a promising overlap area that warrants further research exploration.

Acknowledgments

This research was funded in part by the Indo-French Centre for the Promotion of Advanced Research (IFCPAR/CEFIPRA) through project number CSRP 6702-2.

References

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=MU2495w47rz.

David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, pages 412–421, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1042. URL https://aclanthology.org/D17-1042.

Krishna B. Athreya and Soumen N. Lahiri. *Measure Theory and Probability Theory (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 038732903X.

Clément Bénesse, Fabrice Gamboa, Jean-Michel Loubes, and Thibaut Boissin. Fairness seen as global sensitivity analysis. *Machine Learning*, 113(5):3205–3232, May 2024.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

Nils Ole Breuer, Andreas Sauter, Majid Mohammadi, and Erman Acar. Cage: Causality-aware shapley value for global explanations. In Luca Longo, Sebastian Lapuschkin, and Christin Seifert, editors, *Explainable Artificial Intelligence*, pages 143–162, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-63800-8.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 981–990. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/chattopadhyay19a.html.

Charles Corbière, Marc Lafon, Nicolas Thome, Matthieu Cord, and Patrick Pérez. Beyond First-Order Uncertainty Estimation with Evidential Models for Open-World Recognition. In *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*, Virtual, Austria, September 2021. URL https://cnam.hal.science/hal-03347628.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 203–214, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534159. URL https://doi.org/10.1145/3514094.3534159.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael Emmerich, and Heike Trautmann, editors, *Parallel Problem Solving from Nature – PPSN XVI*, pages 448–469, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58112-1.

Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 915–924, 2022.

Thomas FEL, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26005–26014. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/da94cbeff56cfda50785df477941308b-Paper.pdf.

Nicholas Frosst and Geoffrey E. Hinton. Distilling a neural network into a soft decision tree. *ArXiv*, abs/1711.09784, 2017. URL https://api.semanticscholar.org/CorpusID:3976789.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OPyWRrcjVQw.

David Galles and Judea Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1):9–43, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(97)00047-7. URL https://www.sciencedirect.com/science/article/pii/S0004370297000477. Relevance.

Mathieu Gerber. On integration methods based on scrambled nets of arbitrary size. *Journal of Complexity*, 31(6):798–816, 2015. ISSN 0885-064X. doi: https://doi.org/10.1016/j.jco.2015.06.001. URL https://www.sciencedirect.com/science/article/pii/S0885064X1500059X.

Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *CoRR*, abs/1907.07165, 2019a. URL http://arxiv.org/abs/1907.07165.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/goyal19a.html.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4778–4789. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. ISSN 10618600. URL http://www.jstor.org/stable/27594267.

Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 279–287, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314230. URL https://doi.org/10.1145/3306618.3314230.

Nicholas J. Irons, Meyer Scetbon, Soumik Pal, and Zaid Harchaoui. Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates, 2021. URL https://arxiv.org/abs/2112.15595.

Michiel J.W. Jansen. Analysis of variance designs for model output. *Computer Physics Communications*, 117(1):35–43, 1999. ISSN 0010-4655. doi: https://doi.org/10.1016/S0010-4655(98)00154-4. URL https://www.sciencedirect.com/science/article/pii/S0010465598001544.

Dominik Janzing, Lenon Minorics, and Patrick Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/janzing20a.html.

Dominik Janzing, Patrick Blöbaum, Atalanti A. Mastakouri, Philipp M. Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions, 2024. URL https://arxiv.org/abs/2007.00714.

Adrián Javaloy, Pablo Sanchez Martin, and Isabel Valera. Causal normalizing flows: from theory to practice. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=QIFoCI7ca1.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10476–10501. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jung22a.html.

Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, Vineeth N. Balasubramanian, and Amit Sharma. Matching learned causal effects of neural networks with domain priors. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:246473224.

Ilyes Khemakhem, Diederik P. Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2019. URL https://api.semanticscholar.org/CorpusID:195874364.

Donald E. Knuth and Jayme L. Szwarcfiter. A structured program to generate all topological sorting arrangements. *Information Processing Letters*, 2(6):153–157, 1974. ISSN 0020-0190. doi: https://doi.org/10.1016/0020-0190(74)90001-5. URL https://www.sciencedirect.com/science/article/pii/0020019074900015.

Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 652–663, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462597. URL https://doi.org/10.1145/3461702.3462597.

S. Kucherenko, S. Tarantola, and P. Annoni. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183(4):937–946, 2012. ISSN 0010-4655. doi: https://doi.org/10.1016/j.cpc.2011.12.020. URL https://www.sciencedirect.com/science/article/pii/S0010465511004085.

Julius Von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *2nd Conference on Causal Learning and Reasoning*, 2023. URL https://openreview.net/forum?id=stVikewRRvw.

Sonja Kuhnt and Arkadius Kalka. *Global Sensitivity Analysis for the Interpretation of Machine Learning Algorithms*, pages 155–169. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-031-07155-3_6. URL https://doi.org/10.1007/978-3-031-07155-3_6.

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://api.semanticscholar.org/CorpusID:12533380.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Pierre L'Ecuyer. Randomized quasi-monte carlo: An introduction for practitioners. In Art B. Owen and Peter W. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 29–52, Cham, 2018. Springer International Publishing.

Pierre L'Ecuyer and Christiane Lemieux. *Recent Advances in Randomized Quasi-Monte Carlo Methods*, pages 419–474. Springer US, New York, NY, 2002. ISBN 978-0-306-48102-4. doi: 10.1007/0-306-48102-2_20. URL https://doi.org/10.1007/0-306-48102-2_20.

Genyuan Li, Herschel Rabitz, Paul E. Yelvington, Oluwayemisi O. Oluwole, Fred Bacon, Charles E. Kolb, and Jacqueline Schoendorf. Global sensitivity analysis for systems with independent and/or correlated inputs. *The Journal of Physical Chemistry A*, May 2010.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *ArXiv*, abs/1912.03277, 2019. URL https://api.semanticscholar.org/CorpusID:208857863.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Art B. Owen. Scrambling sobol' and niederreiter–xing points. *Journal of Complexity*, 14(4):466–489, 1998. ISSN 0885-064X. doi: https://doi.org/10.1006/jcom.1998.0487. URL https://www.sciencedirect.com/science/article/pii/S0885064X98904873.

Art B. Owen. Sobol' indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014. doi: 10.1137/130936233. URL https://doi.org/10.1137/130936233.

Art B. Owen and Daniel Rudolf. A strong law of large numbers for scrambled net integration. *SIAM Review*, 63(2):360–372, 2021. doi: 10.1137/20M1320535. URL https://doi.org/10.1137/20M1320535.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6c1da886822c67822bcf3679d04369fa-Paper.pdf.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151. BMVA Press, 2018. URL http://bmvc2018.org/contents/papers/1064.pdf.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3976–3990. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/294e09f267683c7ddc6cc5134a7e68a8-Paper.pdf.

Drago Plečko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024. ISSN 1935-8237. doi: 10.1561/2200000106. URL http://dx.doi.org/10.1561/2200000106.

Arnald Puy, William Becker, Samuele Lo Piano, and Andrea Saltelli. A comprehensive comparison of total-order estimators for global sensitivity analysis. *International Journal for Uncertainty Quantification*, 12(2):1–18, 2022. ISSN 2152-5080.

R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Sharif Rahman. A generalized anova dimensional decomposition for dependent probability measures. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):670–697, 2014. doi: 10.1137/120904378. URL https://doi.org/10.1137/120904378.

Abbaavaram Gowtham Reddy, Saketh Bachu, Harsh Nilesh Pathak, Ben Godfrey, Vineeth N. Balasubramanian, V Varshaneya, and Satya Narayanan Kar. Towards learning and explaining indirect causal effects in neural networks. In *AAAI Conference on Artificial Intelligence*, 2023a. URL https://api.semanticscholar.org/CorpusID:257756923.

Abbavaram Gowtham Reddy, Saketh Bachu, Saloni Dash, Charchit Sharma, Amit Sharma, and Vineeth N Balasubramanian. On counterfactual data augmentation under confounding, 2023b. URL https://arxiv.org/abs/2305.18183.

Carl O. Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Röttger, Heimo Müller, and Andreas Holzinger. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86:101243, 2024. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2024.101243. URL https://www.sciencedirect.com/science/article/pii/S1389041724000378.

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*, KDD '16, page 1135–1144, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

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*, KDD '16, page 1135–1144, New York, NY, USA, 2016b. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Saptarshi Saha and Utpal Garain. On noise abduction for answering counterfactual queries: A practical outlook. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=4FU8Jz1Oyj.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023.

Andrea Saltelli, Ratto Marco, A Terry, Campolongo Francesca, Cariboni Jessica, Gatelli Debora, Saisana Michaela, and Tarantola Stefano. Global sensitivity analysis: The primer. 2008. URL https://api.semanticscholar.org/CorpusID:115957810.

Christian A. Scholbeck, Julia Moosbauer, Giuseppe Casalicchio, Hoshin Gupta, Bernd Bischl, and Christian Heumann. Position paper: Bridging the gap between machine learning and sensitivity analysis, 2024. URL https://arxiv.org/abs/2312.13234.

Patrick Schwab and Walter Karlen. *CXPlain: causal explanations for model interpretation under uncertainty*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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. doi: 10.1109/JPROC.2021.3058954.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

L. S. Shapley. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton, 1953. ISBN 9781400881970. doi: doi:10.1515/9781400881970-018. URL https://doi.org/10.1515/9781400881970-018.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022. URL http://jmlr.org/papers/v23/21-0080.html.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.

Phillip Si, Zeyi Chen, Subham Sekhar Sahoo, Yair Schiff, and Volodymyr Kuleshov. Semi-autoregressive energy flows: exploring likelihood-free training of normalizing flows. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

I.M Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90144-9. URL https://www.sciencedirect.com/science/article/pii/0041555367901449.

I.M Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. ISSN 0378-4754. doi: https://doi.org/10.1016/S0378-4754(00)00270-6. URL https://www.sciencedirect.com/science/article/pii/S0378475400002706. The Second IMACS Seminar on Monte Carlo Methods.

Bas Van Stein, Elena Raponi, Zahra Sadeghi, Niek Bouman, Roeland C. H. J. Van Ham, and Thomas Bäck. A comparison of global sensitivity analysis methods for explainable ai with an application

in genomic prediction. *IEEE Access*, 10:103364–103381, 2022. doi: 10.1109/ACCESS.2022.3210175.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, July 2008.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:16747630.

Andrzej T. Tunkiel, Dan Sui, and Tomasz Wiktorski. Data-driven sensitivity analysis of complex machine learning models: A case study of directional drilling. *Journal of Petroleum Science and Engineering*, 195:107630, 2020. ISSN 0920-4105. doi: https://doi.org/10.1016/j.petrol.2020.107630. URL https://www.sciencedirect.com/science/article/pii/S0920410520306975.

Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 650–665, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86520-7.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review, 2022. URL https://arxiv.org/abs/2010.10596.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018. URL https://arxiv.org/abs/1711.00399.

Yunqi Wang, Furui Liu, Zhitang Chen, Yik-Chung Wu, Jianye Hao, Guangyong Chen, and Pheng-Ann Heng. Contrastive-ace: Domain generalization through alignment of causal mechanisms. *IEEE Transactions on Image Processing*, 32:235–250, 2023. doi: 10.1109/TIP.2022.3227457.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/wimmer23a.html.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent, editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 6912–6939. PMLR, 2023. URL https://proceedings.mlr.press/v206/xi23a.html.

Ankit Yadu, P K Suhas, and Neelam Sinha. Class specific interpretability in cnn using causal analysis. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3702–3706, 2021. doi: 10.1109/ICIP42928.2021.9506118.

Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1563–1570, 2018. doi: 10.1109/HPCC/SmartCity/DSS.2018.00256.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9588–9597, 2021. doi: 10.1109/CVPR46437.2021.00947.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.

Appendix A. Structural Causal Models

A.1. Causality Preliminaries (Kügelgen et al., 2023; Reddy et al., 2023b)

In this Section, we outline the fundamental definitions and concepts necessary to understand our paper.

Definition 8 (Structural Causal Models) *A Structural Causal Model (SCM) $\mathcal{S}(\mathbf{V}, \mathbf{U}, \mathbf{f}, P_{\mathbf{U}})$ represents cause-effect relationships among a set of random variables, divided into endogenous variables $\mathbf{V} = \{V_1, V_2, ..., V_p\}$ and exogenous variables $\mathbf{U} = \{U_1, U_2, ..., U_p\}$, through a collection of structural equations $\mathbf{f} = \{f_1, f_2, ..., f_p\}$. Each variable $V_j \in \mathbf{V}$ is defined in relation to its parents $PA_j \subseteq \mathbf{V} - j$ using the causal law $V_j = f_j(PA_j, U_j)$. $P_{\mathbf{U}}$ is the probability distribution over exogenous variables.*

The causal diagram $\mathfrak{G}(\mathcal{S})$ affiliated with an SCM $\mathcal{S}$ is a directed graph where each node represents a variable, and directed edges point from the elements of PA_j and U_j towards V_j. As exogenous variables $\mathbf{U}$ are typically unobserved, it is common practice to focus only on the subset of $\mathfrak{G}(\mathcal{S})$ projected onto $\mathbf{V}$. A directed graph is acyclic if it contains no cycles, in which case it is called a directed acyclic graph (DAG). A path in a causal graph (DAG) is defined as a sequence of distinct vertices $X_1, X_2, ..., X_n$ such that there is an edge between each pair of consecutive vertices X_i and X_{i+1}. This edge can be either $X_i \to X_{i+1}$ or $X_{i+1} \to X_i$. A directed path is one where all edges point in the same direction. If there is a directed path from X_j to X_i, then X_j is called an ancestor of X_i, and X_i is referred to as a descendant of X_j.

Definition 9 (Interventional Distribution) *The interventional distribution $\mathbf{X}$ under an intervention where X_j is set to a specific value x_j, denoted as $do(X_j = x_j)$, is defined as follows:*

$$P(\mathbf{X}|do(X_j = x_j)) = \mathbf{1}_{X_j = x_j} \times \prod_{i \neq j} P(X_i|PA_i),$$

where $\mathbf{1}$ is the indicator function.

Appendix B. Causal Interpretation

Using the backtracking semantics from Kügelgen et al. (2023), we will now explain the causal interpretation of ICC. "Backtracking" alludes to the process of adjusting upstream variables to account for counterfacts while preserving the underlying causal structures in the system.

Instead of measuring the reduction in uncertainty caused by adjusting the observed value x_j of the node X_j, we assess the reduction achieved by modifying the associated noise u_j. However, through the backtracking, adjustment of a noise u_j can be interpreted as an intervention on X_j without altering the joint distribution of $\mathbf{X}$ (Janzing et al., 2024) : After noting that the parent variables of X_j have taken the values pa_j, we assign X_j the value $x'_j = f_j(pa_j, u'_j)$, where u'_j is randomly sampled from P_{U_j}. As U_j has no parents, we can treat U_j as a randomized treatment. Thus, the interventional probabilities are reduced to observational probabilities

$$P(\cdot|do(U_j = u'_j)) = P(\cdot|U_j = u'_j), \qquad \forall 1 \leq j \leq p.$$

As a result, we did not explicitly write interventional probabilities in the Definition 1. However, a statistical dependence between $\hat{Y}$ and U_j suggests a causal influence of X_j on $\hat{Y}$.

Appendix C. Variance-based Sensitivity Analysis

The Hoeffding decomposition, also known as the ANOVA decomposition or high-dimensional model representation (HDMR), allows us to represent the function f as follows:

$$\hat{Y} = f(\mathbf{X}) = \sum_{T \subseteq [p]} f_T(\mathbf{X}_T), \tag{8}$$

with the functions f_T defined recursively as

$$f_T(\mathbf{X}_T) = \mathbb{E}(f(\mathbf{X})|\mathbf{X}_T) - \sum_{T' \subset T} f_{T'}(\mathbf{X}_{T'}). \tag{9}$$

The ANOVA decomposition satisfies the orthogonality constraint

$$\mathbb{E}_{T \neq T'}(f_T(\mathbf{X}_T) f_{T'}(\mathbf{X}_{T'})) = 0 \quad \forall T, T' \subseteq [p], \tag{10}$$

and let us decompose the model variance as follows:

$$\mathbb{V}(f(\mathbf{X})) = \sum_{T \subseteq [p]} \mathbb{V}(f_T(\mathbf{X}_T)). \tag{11}$$

Definition 10 (Sobol' indices (FEL et al., 2021)) *The sensitivity index $\mathcal{S}_T$, which quantifies the contribution of the variable set $\mathbf{X}_T$ to the model response $f(\mathbf{X})$ in terms of its fluctuations, is defined as:*

$$\mathcal{S}_T = \frac{\mathbb{V}(f_T(\mathbf{X}_T))}{\mathbb{V}(f(\mathbf{X}))}. \tag{12}$$

Sobol indices quantify the proportion of the output's variance caused by any subset of input features. Theorem 1 from Owen (2014) establishes a connection between Sobol' indices and the Shapley value, where the latter is computed using a variance-based value function. We restate this theorem to match our setting.

Theorem 11 *For any $1 \leq j \leq p$,*

$$\sum_{T \subseteq [p]/\{j\}} \frac{1}{p\binom{p-1}{|T|}} \Big(\mathbb{V}(\mathbb{E}(f(\mathbf{X})|\mathbf{X}_{T+j})) - \mathbb{V}(\mathbb{E}(f(\mathbf{X})|\mathbf{X}_T)) \Big) = \sum_{T \subseteq [p], j \in T} \frac{\mathcal{S}_T}{|T|}. \tag{13}$$

Appendix D. Proofs

Lemma 1 (Janzing et al. (2024)) *For a topological ordering $\pi \in \mathcal{C}(\mathcal{G})$, we have*

$$\phi(\hat{Y}|\mathbf{U}_{T_\pi^j}) = \phi(\hat{Y}|\mathbf{X}_{T_\pi^j}) = \phi(\hat{Y}|do(\mathbf{X}_{T_\pi^j})).$$

Proof The first equality holds due to the conditional independence $\hat{Y} \perp\!\!\!\perp \mathbf{X}_{T_\pi^j}|\mathbf{U}_{T_\pi^j}$ and the fact that $\mathbf{X}_{T_\pi^j}$ is function of $\mathbf{U}_{T_\pi^j}$. The second equality is valid as conditioning on all ancestors blocks any backdoor paths. ■

Property 1 (Efficiency/ Completeness)

$$\sum_j ICC_\phi^{To}(X_j \to \hat{Y}) = \phi(\hat{Y}|\mathbf{U}) - \phi(\hat{Y}).$$

Proof

$$\begin{aligned}
\sum_j ICC_\phi^{\text{To}}(X_j \to \hat{Y}) &= \frac{1}{|\mathcal{C}(\mathcal{G})|} \sum_j \sum_{\pi \in \mathcal{C}(\mathcal{G})} ICC_\phi^{\text{To}}(X_j \to \hat{Y}|I = T_\pi^j) \\
&= \frac{1}{|\mathcal{C}(\mathcal{G})|} \sum_{\pi \in \mathcal{C}(\mathcal{G})} \sum_j ICC_\phi^{\text{To}}(X_j \to \hat{Y}|I = T_\pi^j) \\
&= \frac{1}{|\mathcal{C}(\mathcal{G})|} \sum_{\pi \in \mathcal{C}(\mathcal{G})} \underbrace{\sum_j \phi(\hat{Y}|\mathbf{U}_{T_\pi^j \cup \{j\}}) - \phi(\hat{Y}|\mathbf{U}_{T_\pi^j})}_{\phi(\hat{Y}|\mathbf{U}_V) - \phi(\hat{Y})} \\
&= \frac{1}{|\mathcal{C}(\mathcal{G})|} \cdot |\mathcal{C}(\mathcal{G})| \cdot \big(\phi(\hat{Y}|\mathbf{U}_V) - \phi(\hat{Y})\big) = \phi(\hat{Y}|\mathbf{U}) - \phi(\hat{Y})
\end{aligned}$$

■

Property 2 (Nullity/ Dummy)

$$ICC_\phi^{To}(X_j \to \hat{Y}) = ICC_\phi^{Sh}(X_j \to \hat{Y}) = 0$$

whenever $\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}})$ *for all* $I \subseteq [p] - \{j\}$.

Proof

By definition, if $\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}})$, then $ICC_\phi(X_j \to \hat{Y}) = 0$. The result follows immediately when this holds for all $I \subseteq [p] - j$. ■

Property 3 (Symmetry)

$$ICC_\phi^{Sh}(X_j \to \hat{Y}) = ICC_\phi^{Sh}(X_l \to \hat{Y}),$$

if $\phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}}) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{l\}}) \quad \forall I$ *s.t.* $I \subseteq [p] - \{j, l\}$.

Proof

Note that for $I = \emptyset$, we have $\phi(\hat{Y}|U_j) = \phi(\hat{Y}|U_l)$. For any $I \subseteq [p] - \{j, l\}$, if $\phi(\hat{Y}|\mathbf{U}_{I\cup\{j\}}) = \phi(\hat{Y}|\mathbf{U}_{I\cup\{l\}})$, then it follows that $ICC_\phi(X_j \to \hat{Y}|I) = ICC_\phi(X_l \to \hat{Y}|I)$. More importantly, for each $T \subseteq [p] - \{j\}$ with $l \in T$, there exists a corresponding subset $T' \subseteq [p] - \{l\}$ such that $j \in T'$ and

$$\phi(\hat{Y}|\mathbf{U}_T) = \phi(\hat{Y}|\mathbf{U}_{T'}); \quad \phi(\hat{Y}|\mathbf{U}_{T\cup\{j\}}) = \phi(\hat{Y}|\mathbf{U}_{T'\cup\{l\}}).$$

This can be seen from the following:

$$\phi(\hat{Y}|\mathbf{U}_{T\cup\{j\}}) = \phi(\hat{Y}|\mathbf{U}_{T-\{l\}}, U_l, U_j) = \phi(\hat{Y}|\mathbf{U}_{\underbrace{T-\{l\}\cup\{j\}}_{T'}}, U_l) = \phi(\hat{Y}|\mathbf{U}_{T'\cup\{l\}}),$$

and since $T-\{l\} \subseteq [p]-\{l,j\}$, we have:

$$\phi(\hat{Y}|\mathbf{U}_{T'}) = \phi(\hat{Y}|\mathbf{U}_{T-\{l\}}, U_j) = \phi(\hat{Y}|\mathbf{U}_{T-\{l\}}, U_l) = \phi(\hat{Y}|\mathbf{U}_T).$$

As a result, we get

$$ICC_\phi(X_j \to \hat{Y}|T) = ICC_\phi(X_l \to \hat{Y}|T'),$$

with $|T| = |T'|$. From the results above, we can easily deduce the following equality:

$$\begin{aligned}
ICC_\phi^{\text{Sh}}(X_j \to \hat{Y}) &= \sum_{T\subseteq[p]/\{j\}} \frac{1}{n\binom{n-1}{|T|}} ICC_\phi(X_j \to \hat{Y}|T) \\
&= \sum_{T\subseteq[p]-\{j,l\}} \frac{1}{n\binom{n-1}{|T|}} ICC_\phi(X_j \to \hat{Y}|T) + \sum_{\substack{T\subseteq[p]-\{j\} \\ l\in T}} \frac{1}{n\binom{n-1}{|T|}} ICC_\phi(X_j \to \hat{Y}|T) \\
&= \sum_{T\subseteq[p]-\{j,l\}} \frac{1}{n\binom{n-1}{|T|}} ICC_\phi(X_l \to \hat{Y}|T) + \sum_{\substack{T'\subseteq[p]-\{l\} \\ j\in T'}} \frac{1}{n\binom{n-1}{|T'|}} ICC_\phi(X_l \to \hat{Y}|T') \\
&= \sum_{T\subseteq[p]/\{l\}} \frac{1}{n\binom{n-1}{|T|}} ICC_\phi(X_l \to \hat{Y}|T) \\
&= ICC_\phi^{\text{Sh}}(X_l \to \hat{Y})
\end{aligned}$$

■

Property 4 (Sensitivity/ Causal irrelevance) *If X_i is causally irrelevant (Galles and Pearl, 1997) to $\hat{Y}$ for all $I \subseteq [p]-\{i\}$, i.e.,*

$$P(\hat{Y}|do(X_i, \mathbf{X}_I)) = P(\hat{Y}|do(\mathbf{X}_I)), \quad \forall I \subseteq [p]-\{i\},$$

then

$$ICC_\phi^{Sh}(X_i \to \hat{Y}) = ICC_\phi^{To}(X_i \to \hat{Y}) = 0.$$

Proof Note that for $I = \emptyset$,

$$P(\hat{Y}|do(X_i = x_i)) = P(\hat{Y}). \tag{14}$$

The possible (natural) values of $\hat{Y}$ are:

$$\begin{aligned}
\mathcal{Y} &= \{f(x_i', \mathbf{x}_{-i})|x_i' \in \mathcal{X}_i, \mathbf{x}_{-i} \in \mathcal{X}_{-i}\} \\
&= \{f^*(u_i, \mathbf{u}_{-i})|u_i \in \mathcal{U}_i, \mathbf{u}_{-i} \in \mathcal{U}_{-i}\},
\end{aligned}$$

where $\mathcal{X}_{-i}$, $\mathcal{X}_i$, $\mathcal{U}_{-i}$ and $\mathcal{U}_i$ are the supports of $\mathbf{X}_{-i}, X_i, \mathbf{U}_{-i}$ and U_i, respectively. Similarly, under $do(X_i = \mathbf{x}_i)$, the possible values of $\hat{Y}$ are given by:

$$\begin{aligned}\mathcal{Y}_{x_i} &= \{f(x_i, \mathbf{x}_{-i}) | \mathbf{x}_{-i} \in \mathcal{X}_{-i}^{do(x_i)}\} \\ &= \{\tilde{f}(x_i, \mathbf{u}_{-i}) | \mathbf{u}_{-i} \in \mathcal{U}_{-i}\},\end{aligned}$$

where $\mathcal{X}_{-i}^{do(x_i)}$ is the support of $\mathbf{X}_{-i}$ under the intervention $do(X_i = x_i)$. The equality of distributions in equation 14 imposes the constraint that the support of Y under the intervention must match its natural support, i.e., $\mathcal{Y} = \mathcal{Y}_{x_i}$. In other words, for each $(u_i, \mathbf{u}_{-i}) \in \mathcal{U}, \exists \mathbf{u}'_{-i} \in \mathcal{U}_{-i}$ such that $f^*(u_i, \mathbf{u}_{-i}) = \tilde{f}(x_i, \mathbf{u}'_{-i})$. Since $\mathbf{U}$ does not depend on X_i, the function f^* also does not depend on U_i. Therefore, $\hat{Y}$ functionally does not depend on U_i. Thus, for any $I \subseteq [p] - \{i\}$,

$$\phi(\hat{Y}|\mathbf{U}_{i+I}) = \phi(f^*(\mathbf{U}_{-i})|\mathbf{U}_I, U_i) = \phi(f^*(\mathbf{U}_{-i})|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{U}_I).$$

The rest follows directly from Property 2.

■

Theorem 2 *Let X, Y and Z be random variables on the same probability space and $\mathbb{V}(X) < \infty$. Then,*

$$\mathbb{V}_Y(\mathbb{E}(X|Y)) \leq \mathbb{V}_{Y,Z}(\mathbb{E}(X|Y,Z)).$$

Proof The law of total variance states that $\mathbb{V}(X) = \mathbb{E}_Y(\mathbb{V}(X|Y)) + \mathbb{V}_Y(\mathbb{E}(X|Y))$. Similarly, $\mathbb{V}(X) = \mathbb{E}_{Y,Z}(\mathbb{V}(X|Y,Z)) + \mathbb{V}_{Y,Z}(\mathbb{E}(X|Y,Z))$. From this, it follows that the expected variance of X is greater than or equal to the expected value of the conditional variance of X given Y, i.e., $\mathbb{E}(\mathbb{V}(X)) \geq \mathbb{E}_Y(\mathbb{V}(X|Y))$ which also implies the conditional version $\mathbb{E}_Z(\mathbb{V}(X|Z)) \geq \mathbb{E}_{Y,Z}(\mathbb{V}(X|Y,Z))$ for any random variable Z. Interchanging Y and Z in the last expression and subtracting the expected variances from $\mathbb{V}(X)$, we obtain the stated inequality. ■

Theorem 4 *Assume that input features $\{X_i\}$ are independent. Then, with our specific choice of ϕ, for any $I \subseteq [p]$ and any $j \in [p]$, we have*

$$\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{X}_I) = \sum_{T \subseteq I} \mathcal{S}_T; \qquad ICC_\phi(X_j \to \hat{Y}) = \sum_{j \in T, T \subseteq [p]} \frac{\mathcal{S}_T}{|T|}, \tag{7}$$

where $\mathcal{S}_T$ is the Sobol' sensitivity index of the input subset $\mathbf{X}_T$ (see Appendix C for more details).

Proof When the input features are independent, we have $\mathcal{C}(\mathcal{G}) = S_p$. Consequently, from Lemma 1, for any $\pi \in S_p$, the following holds:

$$\phi(\hat{Y}|\mathbf{U}_{T_\pi^j}) = \phi(\hat{Y}|\mathbf{X}_{T_\pi^j}).$$

In other words, for any subset $I \subseteq [p]$, $\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{X}_I)$. For our specific choice of ϕ, we have:

$$\phi(\hat{Y}|\mathbf{U}_I) = \phi(\hat{Y}|\mathbf{X}_I) = \frac{\mathbb{V}_{\mathbf{X}_I}(\mathbb{E}(\hat{Y}|\mathbf{X}_I))}{\mathbb{V}(\hat{Y})} = \frac{\sum_{T \subseteq I} \mathbb{V}(f_T(\mathbf{X}_T))}{\mathbb{V}(f(\mathbf{X})} = \sum_{T \subseteq I} \mathcal{S}_T,$$

where the third equality follows directly from Equations 9 and 10.

With independent input features,it follows from Equations 6 and 5 that ICC_ϕ^{To} and ICC_ϕ^{Sh} coincide:

$$\begin{aligned}
ICC_\phi^{\text{To}}(X_j \to \hat{Y}) &= ICC_\phi^{\text{Sh}}(X_j \to \hat{Y}) \\
&= \sum_{T \subseteq [p]/\{j\}} \frac{1}{p\binom{p-1}{|T|}} ICC_\phi(X_j \to \hat{Y}|T) \\
&= \sum_{T \subseteq [p]/\{j\}} \frac{1}{p\binom{p-1}{|T|}} \Big(\phi(\hat{Y}|\mathbf{U}_{j+T}) - \phi(\hat{Y}|\mathbf{U}_T)\Big) \\
&= \sum_{T \subseteq [p]/\{j\}} \frac{1}{p\binom{p-1}{|T|}} \Big(\phi(\hat{Y}|\mathbf{X}_{j+T}) - \phi(\hat{Y}|\mathbf{X}_T)\Big) \\
&= \sum_{T \subseteq [p]/\{j\}} \frac{1}{p\binom{p-1}{|T|}} \Big(\mathbb{V}(\mathbb{E}(f(\mathbf{X})|\mathbf{X}_{T+j})) - \mathbb{V}(\mathbb{E}(f(\mathbf{X})|\mathbf{X}_T))\Big),
\end{aligned}$$

The rest of the proof is immediate from Theorem 11. ■

Theorem 7 *Under Assumption 2, suppose we have two CNFs $(\mathbf{F}_{\theta_1}, P_{\theta_1})$ and $(\mathbf{F}_{\theta_2}, P_{\theta_2})$ that both match the observational distribution $P(\mathbf{X}, \hat{Y})$, then the intrinsic causal contributions of X_j on $\hat{Y}$ will be equal for both CNFs. Specifically, we have:*

$$ICC_{\phi,\theta_1}(X_j \to \hat{Y}) = ICC_{\phi,\theta_2}(X_j \to \hat{Y}).$$

Proof Continuing from Theorem 6, since h is measurable bijection, the $\sigma-$algebras generated by u_p and $h_p(u_p)$ are identical and thus we have $\mathbb{E}(\hat{Y}|u_p) = \mathbb{E}(\hat{Y}|h_p(u_p))$ (Athreya and Lahiri, 2006). More generally, for any subset $I \subseteq [p]$, it follows that $\mathbb{E}(\hat{Y}|\mathbf{u}_I) = \mathbb{E}(\hat{Y}|h_I(\mathbf{u}_I))$, where $h(\mathbf{u}_I) = h_I(\mathbf{u}_I) = \big(h_j(u_j)\big)_{j\in I}$. Given our specific choice of ϕ, it follows directly that $\phi(\hat{Y}|h(\mathbf{U}_I)) = \phi(\hat{Y}|\mathbf{U}_I)$, thereby establishing the equality

$$ICC_{\phi,\theta}(X_j \to \hat{Y}) = ICC_\phi(X_j \to \hat{Y}). \tag{15}$$

The dependence on θ on the right-hand side arises from the expression $\mathbf{F}_\theta^{-1}(\mathbf{F}(\mathbf{u})) = \mathbf{h}(\mathbf{u})$. Equation 15 states that if a CNF matches the observational distribution, then the ICC computed with respect to the flow does not depend on the flow parameter θ. The statement of the theorem follows immediately as a direct consequence. ■

Appendix E. Algorithms

Algorithm 2 Pseudocode for estimating ϕ for ICC^{To}

Input: Batch size B, context I for conditioning, the neural network $\mathcal{N}$

1: $\mathbf{x}_M^{(i)}, \mathbf{x}_N^{(i)} \sim D$ for $i = 1, 2, ..., B$, where D is the dataset.

2: $\mathbf{x}_Q^{(i)} = \left(\mathbf{x}_{M_{-I}}^{(i)}, \mathbf{x}_{N_I}^{(i)}\right)$ for $i = 1, 2, ..., B$.

3: $\hat{y}_M^{(i)} = \mathcal{N}\left(\mathbf{x}_M^{(i)}\right), \hat{y}_N^{(i)} = \mathcal{N}\left(\mathbf{x}_N^{(i)}\right), \hat{y}_Q^{(i)} = \mathcal{N}\left(\mathbf{x}_Q^{(i)}\right)$ for $i = 1, 2, ..., B$.

4: $\bar{y} = \frac{1}{2B} \sum_{i=1}^{B} \left(\hat{y}_M^{(i)} + \hat{y}_N^{(i)}\right); V = \frac{1}{2B-1} \sum_{i=1}^{B} \left((\hat{y}_M^{(i)} - \bar{y})^2 + (\hat{y}_N^{(i)} - \bar{y})^2\right)$

5: $\hat{\psi} = V - \frac{1}{2B} \sum_{i=1}^{B} \left(\hat{y}_N^{(i)} - \hat{y}_Q^{(i)}\right)^2$

Output: $\hat{\phi} = \frac{\hat{\psi}}{V}$

Appendix F. Experiment Setup and Datasets

We usually train the neural networks $\mathcal{N}$ and NFs for 100 epochs with a batch size of 32, using the Adam optimizer with a learning rate of 3×10^{-4}. Only for COMPAS dataset, we train the neural network using a learning rate of 10^{-3}.

Synthetic Data We employ the same data generation process as Reddy et al. (2023a) for the synthetic data experiment. Figure 2b contains the causal graph and the detailed specification of the SCM. In this dataset, the input features W, Z, and X are connected through linear equations with additive Gaussian noise. The output Y is a non-linear function of these inputs, also incorporating additive Gaussian noise. The training set consists of 700 samples, while the test set contains 300 samples.

Auto-MPG We use the Auto-MPG dataset to predict miles per gallon (MPG) based on features including the number of cylinders (C), displacement (D), weight (W), horsepower (H), acceleration (A), and miles per gallon (M). The ground truth causal graph for Auto-MPG is unknown, so we adopt the causal graph proposed by Reddy et al. (2023a), shown in Figure 2c. This graph is constructed using relevant domain knowledge and validated through consultations with GPT-3.5 (Brown et al., 2020) to confirm the correctness of each causal edge. The training set includes 274 samples, and the test set includes 118 samples.

COMPAS The dataset comprises criminal records and demographic features for 6,172 defendants who were released on bail in U.S. state courts between 1990 and 2009. The objective herein is to classify each defendant into one of two categories: bail (indicating they are unlikely to commit a violent crime if released) or no bail (indicating they are likely to commit a violent crime). The causal graph in Figure 2a for the COMPAS dataset is inspired by Plečko and Bareinboim (2024). The training set comprises 4,937 samples, while the test set comprises 1,235 samples.

Appendix G. Comparison with Janzing et al. (2024)

The fundamental difference between their work and ours is that their work does not aim to explain a downstream pre-trained model (neural networks, in our case), whereas this is the primary objec-

tive of our study. While Janzing et al. (2024) introduce the notion of ICC, they do not provide a clear, general algorithm for its estimation. Instead, their experimental setup relies on restrictive assumptions — such as treating an additive noise model as a convenient approximation of structural equations (AutoMPG), or inferring the SCM (River flows) from common-sense knowledge with all regression coefficients set to 1. Moreover, they do not discuss identifiability. In contrast, our framework, which incorporates causal normalizing flows, is more general and ensures identifiability.

Janzing et al. (2024) utilized the publicly available ICC implementation in DoWhy (Blöbaum et al., 2024). Specifically, the `gcm.intrinsic_causal_influence` function was used with the auto-assign feature. The `gcm.intrinsic_causal_influence` function returns ICC within a causal model but is not designed to generate explanations for a pre-trained neural network. For each node in the causal graph, `gcm.auto.assign_causal_mechanisms` fits various regression or classification models and selects the optimal one. In contrast, we model the entire causal data-generating process of the inputs using a single deep neural network (Javaloy et al., 2023).